Statistics for Sociologists

HAGOOD   PRICE

Revised Edition

Statistics for Sociologists

STATISTICS

for

SOCIOLOGISTS

REVISED EDITION

MARGARET JARMAN HAGOOD

United States Department of Agriculture

DANIEL O. PRICE

University of North Carolina

HENRY HOLT AND COMPANY

NEW YORK

COPYRIGHT, 1941, BY MARGARET JARMAN HAGOOD

COPYRIGHT, 1952, BY HENRY HOLT AND COMPANY, INC.

LIBRARY OF CONGRESS CATALOG NUMBER 52–7026

PRINTED IN THE UNITED STATES OF AMERICA

Preface to Revised Edition

The general organization of material into parts and chapters has been retained as in the first edition with exceptions to be noted. All chapters have been reviewed and modified when necessary, especially in three respects: (1) to include developments in methods or applications that have occurred in the last eleven years; (2) to use more current materials for illustrations; (3) to reduce space devoted to computation procedures.

Part V has been omitted. This specialized field cannot be adequately treated in a general text in social statistics and is to be covered in a specialized text being written by one of the authors and others.

New chapters have been added on sampling in social surveys and on applications of factor or component analysis in social research; also the chapter on indexes and scales has been completely rewritten to include Guttman's contributions. Some of the more frequently used formulas have been collected in an Appendix for easy reference. Supplementary readings have been revised to include references to new material and to revised editions of works cited in the first edition. We regret that the progress in application of statistics to sociological research since 1941 has not provided a basis for a more complete substitution of new titles.

The work of making the revision has been done mainly by the junior author, but the revised edition is a product of both authors, who share responsibility for merits or defects.

Washington, D. C.   M. J. H.
Chapel Hill, N. C.   D. O. P.
May 26, 1952
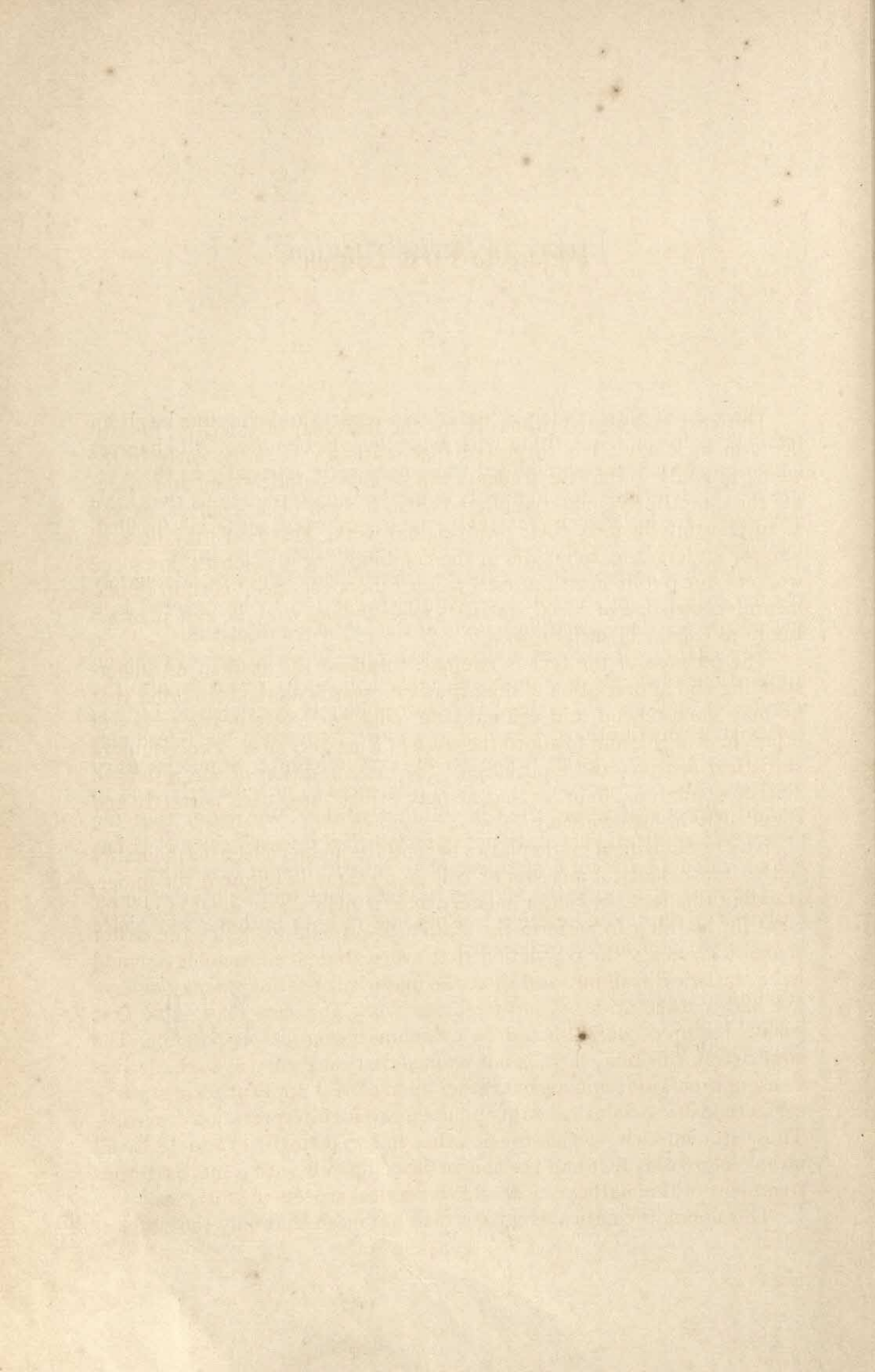
Preface to First Edition

This text is designed primarily for the first year of statistics for students in sociology. It is suggested that statistics be taken early in the course of study so that the students will be able to understand and evaluate the quantitative material offered them in their other courses, as well as to plan intelligently their own original work. This text may be used also for undergraduate majors in the combined social sciences, for social workers, for public health nurses, or for others who need certain fundamental knowledge of social statistics even though they do not plan actually to engage in social research.

The purpose of the text is twofold: to afford the student an understanding and appreciation of quantitative research methods in order that he may comprehend and evaluate the past and current researches of others in his field; and to afford the student a mastery of the procedures of statistical analysis and especially of the interpretation of the results of statistical analysis, in order that he may utilize these tools correctly and confidently in his own work.

No prerequisite of mathematics beyond freshman college mathematics (either mathematical analysis or college algebra) is required for understanding this text. In fact, a mastery of only high school algebra is sufficient for learning to perform the statistical procedures. Since the writer shares with many the conviction that every student in sociology should have statistical training, and since the majority of sociology majors have not had mathematics beyond freshman work, it seems reasonable that such a required course should be as nonmathematical as possible. The emphasis in this text, then, is not upon statistical theory as such, derivations, or proofs of formulas, but rather upon correct application of statistical methods to sociological data and upon careful interpretation of results. Those students who decide to specialize in social statistics should by all means follow this text and the course based upon it with a more rigorous treatment of the mathematical and theoretical aspects of statistics.

This nonmathematical emphasis does not mean that only the simplest

elementary topics are treated here. On the contrary, we have attempted to include all the basic statistical methods which have been used in sociological research, and some of the newer ones which have not as yet found wide application in sociological fields. Furthermore, there is considerable attention to measures of precision, confidence limits, and special methods and formulas for small samples, which are not found in any of the existing texts on social statistics. The writer believes that these more accurate modern statistical methods must supersede some of the looser ones now in use, and that in spite of their mathematically difficult basis, they can be mastered and correctly applied by students who have little mathematical background.

An introduction to statistics which utilizes data, applications, and interpretations in one's own field is undoubtedly easier and more efficient than the alternative of learning methods as applied in another field and then transferring the procedures to one's own. Therefore, this text, written primarily for sociology students, uses materials dealt with by sociologists and those in closely related fields, such as social work, and its focus is on the applications of the various methods to the problems of sociology.

Sociology lags not only behind the physical sciences, but also behind psychology, education, and economics in the application of quantitative methods of research to its problems. Because of the present limitations of the application of statistical methods to sociology, it is possible for the sociology student to master in a year the methods generally utilized in sociology and presented in this text. These same limitations, however, offer a challenge to oncoming sociologists to extend to their own domain the range of application of tools already available.

Most of the material in Parts I, II, and III is covered in a one-quarter introductory course given by the writer, while that of Part IV is covered in a second course. Other teachers who wish to include correlation in a first course may prefer omitting parts or all of Part III and going directly to Chapter 23 from Part II.

Acknowledgments

Contents

Part IV STATISTICS OF RELATIONSHIP

PART I
Quantitative Methods in Sociology
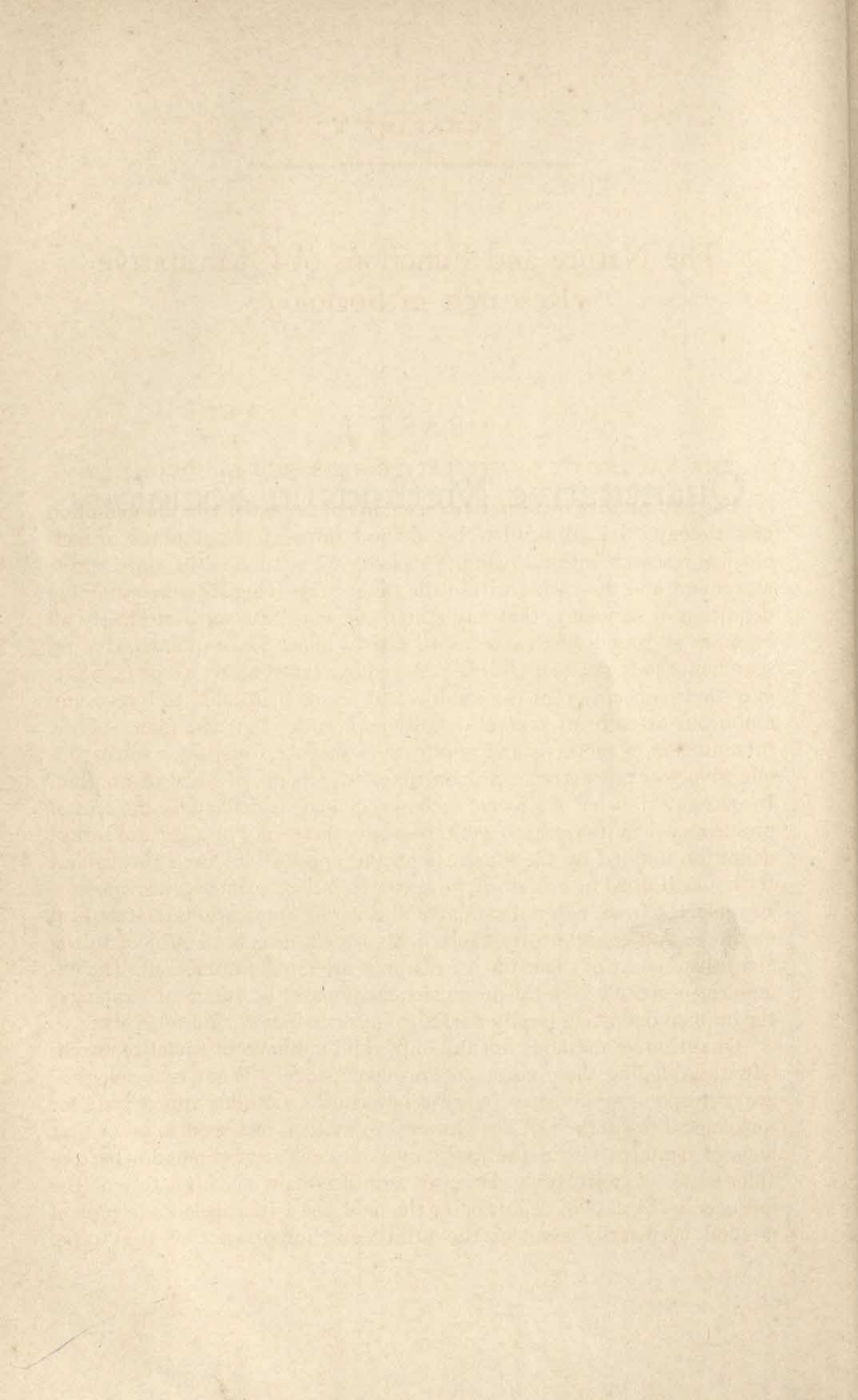
~~~~~~

The Nature and Functions of Quantitative Research in Sociology

THE NATURE OF QUANTITATIVE RESEARCH IN SOCIOLOGY

Use and value of quantitative research not dependent on the definition of sociology. It is difficult to begin the treatment of one phase of sociological research without defining sociology. Yet such is the state of disagreement and dissension within the ranks of sociologists concerning the definition of sociology, that any statement which attempts to satisfy all becomes so broad and vague as to satisfy none. Since quantitative research methods can be usefully employed by almost all sociologists, whatever their conception of their subject, it seems justifiable to forego any ambitious attempt at general definitions in order that the more specific presentation of methods and applications may be found in a setting inoffensive to even extreme proponents of one point of view or another. Insofar as a division of science, such as sociology, is defined by the area of problems which its research workers study, there may be a certain sort of definition implied by the illustrations and applications used throughout the book. It must be explained, however, that these examples are of necessity selected from material available illustrating applications of statistical methods, and consequently those fields which have been subject to the greatest amount of quantitative research are most emphasized. The examples presented here by no means encompass the range of sociology; the implied definition is only partially suggestive, not comprehensive.

Quantitative methods not the only valid methods of social research. Often paralleling the divisions of sociologists over "What is sociology?" are corresponding divisions over "What are the methods appropriate for sociological research?" Of late, interest in method has been so great that lines of demarcation in the ranks over method are overshadowing the older issues of controversy. Here we cannot remain so completely on the fence as in the case of definition of the field, for a treatise on one type of method necessarily assumes the utility and importance of that type.

3

However, we make no all embracing claims for the quantitative as the only type of method useful and appropriate for social research. Our position is simply that quantitative methods are useful in the investigation of many sorts of sociological phenomena, and that when they are mastered more thoroughly by those engaged in sociological research, they will be extended to other areas which have not yet yielded to measurement.

Preliminary definitions of quantitative research methods and of statistics. Quantitative methods of research are those which utilize enumeration and measurement, direct or indirect, relatively accurate or roughly approximate. Statistics (used with a singular verb) is the body of methods developed to deal with data secured through enumeration and measurement. Such data, either raw or condensed, are often called "statistics" (used with a plural verb). Although the origin of the word "statistics" has its basis in the forerunner of modern social statistics, the body of statistical theory has been developed by mathematicians, geneticists, psychologists, economists, and agricultural biologists more than by sociologists. Helen M. Walker's *Studies in the History of Statistical Method with Special Reference to Certain Educational Problems* [1] is an interesting and adequate account of the historical development of statistics; we refer the reader to it for such information. It is sufficient here to state that a body of reliable methods has been developed and is available to the sociologists for their quantitative research.

THE FUNCTIONS OF STATISTICS IN SOCIOLOGICAL RESEARCH

Our chief interest is not in statistical theory for itself, but in what statistics can offer to sociology—and specifically, in the functions of statistical analysis in sociological research. The parts of this book correspond to a division of the topics of statistical analysis on the basis of these several functions. We shall outline the functions and then examine each with an illustration of it from previously published sociological research.

Divisions of statistical methods according to function. The whole body of statistical methods can be divided into two parts: (1) the *descriptive* methods, which condense and summarize data on enumerable or measurable characteristics of a series of units and thus afford a description of certain aspects of those units; (2) the *inductive* or *generalizing* methods, which give estimates of the distributions of characteristics in the larger population or universe from which a sample of units has been drawn and studied. Another twofold division of statistical methods can be made according to complexity: (1) the *simple* methods, which treat single distri-

---

[1] *Studies in the History of Statistical Method with Special Reference to Certain Educational Problems* (Baltimore: Williams and Wilkins, 1929).

butions of characteristics; (2) the *complex* methods which treat distributions of two or more characteristics simultaneously, thus revealing information on the association, or relationship, between two or more characteristics. Hence, the *complex* methods may be called *statistics of relationship*.

These two twofold divisions divide the body of statistical methods into four parts as shown in the outline below.

OUTLINE OF STATISTICAL METHODS ACCORDING TO FUNCTIONS

	Simple Methods (For treating single distributions of characteristics)	*Complex Methods* (For treating two or more distributions of characteristics simultaneously)
Descriptive Methods	I. SIMPLE DESCRIPTIVE STATISTICS	III. DESCRIPTIVE STATISTICS OF RELATIONSHIP
Inductive Methods	II. SIMPLE INDUCTIVE STATISTICS	IV. INDUCTIVE STATISTICS OF RELATIONSHIP

The four divisions of statistics designated in the outline by the four Roman numerals are admitted to be somewhat arbitrary, but they seem to be useful for organizing and explaining statistical methods and therefore will be employed in this book. The names of these divisions will be shortened for convenience, since this can be done without confusion. Thus we shall refer to the four divisions of statistical methods as the following:

I. DESCRIPTIVE STATISTICS
II. INDUCTIVE STATISTICS
III. DESCRIPTIVE STATISTICS OF RELATIONSHIP
IV. INDUCTIVE STATISTICS OF RELATIONSHIP

Often divisions III and IV will be referred to together as STATISTICS OF RELATIONSHIP.

Descriptive statistics. The first function of statistics in the service of research is to afford condensed and summarized descriptions of units with regard to enumerable or measurable characteristics. We call the statistical methods by which this function is achieved *simple descriptive statistics* or, in shortened form, *descriptive statistics*. The measures usually used in descriptive statistics are ratios, rates, percentages, frequency distributions, measures of central tendency (averages), and measures of dispersion (variation of items from averages). These summarizing measures condense masses of unwieldy data into forms which supply information efficiently. Often detailed data on numerous items may be distilled by statistical analysis into one or two summarizing measures which retain

all the essential information. This makes it possible to comprehend more easily certain aspects of the material and also to convey the information to others easily.

An extension of the simplest descriptive methods called a time series is the description of the distribution of a characteristic at different times for the same unit. Another extension is the construction of compound, indirect measures called index numbers. In all of these relatively simple descriptive functions statistics serves to summarize quantitative data about sociological subject matter.

Example of the use of descriptive statistics. The simple descriptive function of statistics may be illustrated by an example from *Understanding Society: The Principles of Dynamic Sociology* by Howard W. Odum.[2] On pages 53 and 54 there are a table, a map, and a bar diagram which together give a summary description of the distribution of people aged 15 to 24 years in the United States in 1940. On the original schedules of the 1940 Census of Population, from which these data are taken, the age of each individual is recorded. If one were to examine all of these original schedules, however, he would not get a very clear picture of the distribution of people aged 15 to 24 because there were more than 130,-000,000 people in the United States at that time, and this is too many numbers for the human mind to comprehend simultaneously. However, the presentation of these figures in tabular form makes them more comprehensible. We can see the number of people of this age in each state and what percentage this number is of the total state population. If these percentages are grouped into classes and shown on a map or presented in a bar chart, we get a concise summary of the distribution of people of this age in the United States. The minute details, such as the age of each individual, have been sacrificed, but we have a comprehensible result.

It may be surprising to some to know that a great amount of quantitative research in sociology involves no more elaborate mathematical treatment than that illustrated above. It is true that some of the theoretical statisticians do not wish to dignify the type of methods which merely summarize and condense quantitative data with the term "statistics," preferring to reserve that term for inductive methods involving the concept of probability. But, in actual practice, so much of the work that goes under the name of statistical research is of this sort that it seems valid to designate such methods as statistical methods, differentiating them from the inductive methods by the label "descriptive." And it may be encouraging to some to know that, up to the present at least, probably a substantial majority of the applications of statistical methods to sociology have utilized only the relatively easy methods of descriptive statistics.

[2] *Understanding Society: The Principles of Dynamic Sociology* (New York: Macmillan, 1947), pp. 53–54.

Inductive statistics. Because of practical considerations a scientific inquiry must often be limited to the study of only a small fraction of the items in which we are interested. When this is the case, a limited number of units—*a sample*—is chosen from the entire series—the *universe*. The method of choosing a sample for study must be carefully planned in order that the sample will be representative of the universe with respect to the characteristic or characteristics being investigated. For analyzing data gathered from a representative sample of units, we have, in addition to the descriptive statistics which afford a summary description of the sample itself, a more elaborate body of methods which permits us to make certain estimates and to draw certain conclusions about the larger group from which the sample has been drawn. The statistical methods used in this inductive function may be called *inductive statistics*. The methods grouped under this name make it possible for us to compute for the whole series of phenomena, that is, for the universe, estimates of the same summarizing measures which the methods of simple descriptive statistics enable us to compute for the sample—ratios, rates, measures of central tendency and dispersion, and other measures summarizing the distribution of a single characteristic.

When we are making generalizations about a universe from which a sample has been drawn and studied, we cannot determine the value of the summarizing measures with the same degree of precision that we can for the sample studied. We can only make estimates of the universe or population values of such summarizing measures. If the method of selecting the sample has been random, we are able not only to make an estimate of the universe value of the summarizing measure, but also to make an estimate of the precision of the first estimate. We do this by designating two values, one above and one below the estimated value, which we can say enclose the universe value and have confidence that in the long run we will be right for whatever percentage of times we choose—usually 95 or 99 percent. Such procedures lead to consideration of matters which cannot be explained at this stage—probability, the normal curve, measures of "error," levels of significance, and confidence or fiducial limits and intervals. We can point out, however, that in the past 40 years there has been great advance in the theory and application of the statistical methods developed for this function of statistics. This advance has been especially marked in the development of treating small samples; no longer can the blanket condemnation of "generalizing from too small a sample" be applied if the research worker has taken careful account of these developments and applied them in drawing his conclusions.

Example of the use of inductive statistics. A notable example of the use of this group of methods is in the Census Bureau's monthly population sample survey. In this sample about 25,000 households located in 68

areas in 42 states and the District of Columbia are interviewed each month. This sample is chosen so that it will be representative of the population of the United States. From the data gathered many totals and percentages are computed by the methods of descriptive statistics to describe the sample. Next, the methods of inductive statistics are employed to make estimates of some of the corresponding totals and percentages for the United States as a whole. For example, we find information on "Marital Status and Household Characteristics: March 1950" in the Current Population Reports, Population Characteristics, Series P–20, No. 33 (February 12, 1951) accompanied by the following statement:

Since the estimates . . . are based on sample data, they are subject to sampling variability. The following table presents the approximate sampling variability of estimates of selected sizes for over-all totals, i.e., those not classified by the items noted below for 1950. The chances are about 19 out of 20 that the difference between the estimate and the figure which would have been obtained from a complete census is less than the sampling variability indicated below:

Size of estimate	Sampling variability	Size of estimate	Sampling variability
10,000	11,000	3,000,000	190,000
50,000	25,000	5,000,000	240,000
100,000	35,000	10,000,000	330,000
300,000	60,000	20,000,000	440,000
500,000	78,000	40,000,000	560,000
1,000,000	110,000		

Estimates of characteristics by urban and rural residence are subject to somewhat greater sampling variability, and estimates of characteristics by age and sex are subject to slightly less sampling variability than that shown above.

The extensive information published in the Current Population Reports of the Bureau of the Census is evidence of the value of inductive statistics.

Descriptive and inductive statistics of relationship. The description of the incidence or distribution of a single characteristic, which both simple descriptive and simple inductive statistics afford, is not sufficient for the purposes of the inquiry in many cases. Another function of statistics is to show the association between two or more characteristics on which we have series of measurements or enumerations. If for any one group of units we have two or more sets of enumerations or measurements, there are methods by which we can ascertain and concisely describe the existence, direction, degree, and nature of the association between the two or more characteristics enumerated or measured. So long as we limit our conclusions and interpretations to the units studied, we

employ *descriptive statistics of relationship* (sometimes called historical statistics of relationship). When we generalize about the several aspects of association in the universe from which the sample has been drawn, we employ *inductive statistics of relationship*. For the former we can determine summarizing measures from data on the series of units studied, and for the latter we can make estimates of the universe values of these summarizing measures. Some of the summarizing measures used in the description of association are coefficients of contingency, coefficients of correlation—total, partial, and multiple; linear and curvilinear—and their corresponding coefficients of regression. Analysis of variance and analysis of covariance are methods for analysis of relationship between different combinations of quantitative and nonquantitative characteristics. As in the case of the inductive statistics for treating one distribution, there are in the inductive statistics of relationship measures of precision and methods of setting confidence limits to the estimated universe values of the coefficients describing the association.

The formulation of the description of the existence, direction, degree, and nature of the association between two or more characteristics in a carefully defined universe (specified as to time, place, and other characteristics) is regarded by some as the equivalent in sociological research to the statement of the "laws" of the physical sciences. Often supplementary, nonstatistical knowledge can lead to a valid interpretation of such association as cause and effect relationship, although statistics per se can never indicate or prove cause and effect. At this stage, which is the highest reach of quantitative sociological research, only those who have an extensive and intimate knowledge of the subject matter and who have also a thorough knowledge and mastery of the statistical methods used are qualified to make the interpretations. Therefore, an expert sociologist who does not know statistics should not farm out the statistical analysis on a research project and then himself draw conclusions on the basis of the results whose potentialities and limitations he cannot fully understand; nor should the most expert statistician who does not know the field from which data are drawn try to interpret the results of his analysis of the data.

The writers wish to reiterate this thesis for emphasis. For sociological research, statistics is a tool extremely valuable to those who have mastered its use, extremely dangerous to those who have not. On the other hand, a knowledge of statistics alone does not qualify one for research in any field save that of statistical theory. Therefore, both a knowledge of sociology and a mastery of statistics are necessary qualifications for quantitative sociological research. While this text is concerned with supplying only the latter qualification (together with the application and interpretation of statistical methods in the field of sociology), it is as-

sumed that any curriculum for training in social research includes other content courses which provide the former.

Example of the use of statistics of relationship. An example of the use of historical or descriptive statistics of relationship may be found in an article by Calvin F. Schmid, "Generalizations Concerning the Ecology of the American City." [3] The article reports an investigation of the relationships between several variables related to the ecology of cities with coefficients of correlation, correlation ratios, and regression equations used as measures of existence, direction, degree, and nature of association. The data for this study were taken from the figures on census tracts in the 1940 census. Measures of association were computed for each of 20 cities.

As an illustration of the procedure in a specific case, let us trace the investigation of the association between mean monthly rental and median school grade completed in the city of Columbus, Ohio. Measures of each of these variables were obtained for each census tract in Columbus. Then a coefficient of correlation between mean monthly rental and median school grade completed was computed. A full explanation of the technique of computation and interpretation of correlation analysis will not be given until Part IV, but some idea of the meaning of a coefficient of correlation can be grasped at this stage. The coefficient of correlation between rentals and school grade completed for Columbus was found to be +.87. Let us see what information this conveys. First we must keep in mind the fact of the *variation* of the 61 census tracts in each of the measures. "Association" refers to the aspect of *covariation* in two measures, their varying together rather than independently. The coefficient of correlation gives information on the existence, direction, and degree of association. Since .87 is different from zero, it informs us as to the *existence* of association—that these two factors are associated. Since .87 is positive, it informs us as to the *direction* of association—that these two variables are *directly* associated, which means that on the average the tracts which have high median school grade completed also have high average rents. Since .87 is closer to 1.00 than to zero, it informs us as to the *degree* of the association—that these two variables are *closely* associated. A coefficient of 1.00 means perfect direct association. The square of the coefficient of correlation, $(.87)^2$ or .76, indicates the proportion of variation in one factor which is associated with the variation in the other factor. In other words, 76 percent of the variation of the 61 census tracts in mean rental is associated with the variation of the tracts in median school grade completed. (Note that we cannot say that 76 percent of the variation in rents is caused by the variation in median school grade com-

---

[3] *American Sociological Review*, 15 (April 1950), pp. 264–281.

pleted.) The regression equation $Y_c = -28.55 + 6.02X$ gives informa-tion on the *nature* of the association but we will leave the discussion of this until Part IV.

When findings such as these are not generalized to any population different from the one actually studied, we call them "descriptive" or "historical" statistics of relationship. Since the case of inductive sta-tistics of relationship is not only more complex but also involves contro-versial principles in interpretation, we omit an example of it here.

Organization of the book in relation to these functions. Part I of this text is introductory in nature. Parts II, III, and IV comprise re-spectively the statistical methods which accomplish these four functions, with the last two treated together.

At this stage it has not been possible to do more than roughly sum-marize the functions of statistics in sociological research. Fuller exposition and discussion of the several functions must be delayed until there have been introduced some of the actual methods and language of statistics. For example, consideration of types of reasoning as related to the different philosophies of science, examination of the nature and limitations of in-ductions drawn from sociological observation and experimentation, and differentiation between generalizing to a limited existent universe and to an infinite hypothetical universe must wait until after the subject of probability has been dealt with. Similarly, the particular advantage to sociological research of certain correlation techniques, which statistically hold certain uncontrollable variables constant while the effect of others is investigated, cannot be intelligently discussed until the methods of correlation analysis have been presented.

It is hoped that enough of the uses of statistics in sociological research have been cursorily indicated to make the prospective student of sta-tistical methods for sociological research aware of the tremendous po-tentialities of this powerful tool of research. Although prediction is the ultimate aim of sociology, certainly it must be preceded by knowledge. A tool which offers so much flexibility and precision in advancing sci-entific knowledge as statistics does should unquestionably be mastered and utilized by those who are seeking to push back the boundaries of the pitifully circumscribed area of valid sociological knowledge. For such knowledge is a necessary condition to the prediction and eventual mastery of societal phenomena.

SUGGESTED READINGS

Bernard, L. L., *The Fields and Methods of Sociology* (New York: Farrar and Rinehart, 1939).

Chapin, F. Stuart, *Experimental Designs in Sociological Research* (New York: Harper, 1947).

Fisher, R. A., *The Design of Experiments*, 5th ed. (New York: Hafner, 1949), Chap. 1.

Johnson, Palmer O., *Statistical Methods in Research* (New York: Prentice-Hall, 1949), Chap. 1.

McCormick, Thomas C., *Elementary Social Statistics* (New York: McGraw-Hill, 1941), Chap. 1.

Odum, Howard W., and Jocher, Katharine, *An Introduction to Social Research* (New York: Holt, 1929), Chap. 18.

Rice, Stuart A. (ed.), *Statistics in Social Studies* (Philadelphia: University of Pennsylvania Press, 1930), Chap. 1.

Stephan, Frederick F., "History of the Uses of Modern Sampling Procedures," *Journal of the American Statistical Association*, 43 (March 1948), pp. 12–40.

Stouffer, Samuel A. and others, *Measurement and Prediction*, Studies in Social Psychology in World War II, Vol. IV (Princeton: Princeton University Press, 1950), Chap. 1.

Taylor, Carl C., "The Social Survey and the Science of Sociology," *The American Journal of Sociology*, 25 (May 1920), pp. 731–756.

Tippett, L. H. C., *Statistics* (London and New York: Oxford University Press, 1943).

Walker, Helen M. "Statistical Literacy in the Social Sciences," *The American Statistician*, Vol. 5 (February 1951), pp. 6–12.

Wirth, Louis (ed.), *Eleven Twenty-Six: A Decade of Social Science Research* (Chicago: University of Chicago Press, 1940).

The Plan and Execution of a Quantitative Research Project

The growth of scientific sociology. The conquest of the wide expanse of sociological ignorance will be brought about by different sorts of contributions. One type of contribution will undoubtedly be the small increments of verifiable knowledge resulting from carefully designed quantitative research projects. When a new bit is added to the accumulation of the previous results of others, it should be ordered into a consistent relation with the body of currently accepted "knowledge." If this is impossible, there must be suspension of judgment until further research can validate either the new or the old, or until some new point of view or unifying theory is developed which provides a new basis of consistency for the apparently opposing sets of "facts."

The place of quantitative research projects in the growth of a science. In the process of growth of a science there are required different sorts of thought, study, and research. Those who have best mastered the existing body of knowledge of a field are best fitted to synthesize what is known and to point out the areas needing further work. For example, P. K. Whelpton in *Needed Population Research*,[1] Rupert B. Vance in *Research Memorandum on Population Redistribution within the United States*,[2] and Dale Yoder in *Demands for Labor: Opportunities for Research*[3] do this for their respective fields and even formulate specific questions which need to be investigated. A research worker who is interested in some such problem, who finds a way of collecting new relevant data or of analyzing in a new way data already collected, and who thus provides new light on the problem goes through the steps of planning and executing a research project. When his contribution has been made available to

---

[1] *Needed Population Research* (Lancaster, Pa.: Science Press, 1938).
[2] *Research Memorandum on Population Redistribution within the United States* (New York: Social Science Research Council, Bulletin 42, 1938).
[3] *Demands for Labor: Opportunities for Research* (New York: Social Science Research Council, Pamphlet 7, 1948).

others, then synthesizers, systematizers, or other theorists may use the contribution in performing their function in the development of the science.

The types of reasoning and the ways by which knowledge is gained are varied. It may be, as some contend, that new knowledge comes from hunches and intuitive insight and is only confirmed by quantitative research. Or it may be, as others claim, that all new knowledge comes through the process of induction—that the collection, analysis, and interpretation of quantitative data is the creative phase in science. The different points of view as to the functions of quantitative research in the development of a science will be treated more fully later; they are mentioned here only to show that quantitative research has a definite place in the procedures outlined by proponents of either side of the controversy. In either case, the quantitative research must be done in units, which we shall call quantitative research projects. Whatever may precede or follow the quantitative project is of great interest to the student of social statistics, but the prerequisite mastery of a specific field and the subsequent placing of the interpretation of the results of a project into a system of sociological knowledge are left for content courses and texts and for more general courses and texts on social research.

STEPS IN PLANNING AND EXECUTING A RESEARCH PROJECT

Since in actual practice planning and executing a specific research project overlap in time sequence, we do not separate these phases artificially, although, of course, as much as possible of the planning precedes the execution. We do, however, for convenience in treatment divide the process of planning and executing a quantitative research project into seven steps or stages and attempt to explain these one at a time. These steps also overlap and sometimes their order may be inverted, but they are presented in their most usual sequence.

1. Formulating the specific questions of investigation in relation to the general problem or field of inquiry. Often this step, listed as the first, is accomplished only after most of the other steps have been completed. Yet, ideally, it should precede the others. One should know what he is trying to find out before he sets about trying to find it. It is here that students who are inexperienced in research frequently fail to narrow and focus their efforts to achievable units. When thesis subjects as broad as juvenile delinquency in the South or differential fertility in the United States are chosen—subjects which transcend any bounds of possibility of accomplishment during graduate work and which are not definitely formulated—they are likely to bring the young research person to a state

of despair when he realizes that the masses of material he has assembled answer no questions, neither confirm nor refute any hypotheses, and yield nothing toward developing a scientific sociology. This step is especially difficult for the very ambitious student because he does not wish to limit his endeavors to the answering of one or two questions which may seem of minor importance to him. It must be urged that the student not take this point of view; the beginning social research worker can make most certain the value of his contribution if he narrows his research to very specific problems. By these bits scientific knowledge grows, and by revealing these bits the student learns not only the importance and full meaning of the knowledge itself, but also the valid methods of acquiring knowledge. Thus he trains himself eventually to tackle larger problems and to gain insight into the underlying principles by which these bits may be synthesized.

The sort of narrowing that must be done consists in formulating such specific questions as, "What is the relation between the incidence of juvenile delinquency, in a certain locality, during a certain period of time and some other factors, such as income of families, percentage of families on relief, or playground facilities?" A part of the formulation consists of precisely defining terms. This must be done not abstractly but in accordance with the concrete aspects of the data available. Therefore, this step must overlap the next.

2. Developing a method of obtaining facts which will answer or at least throw light upon the questions formulated. This step calls for the exercise of ingenuity and constructive creative effort, whether its accomplishment involves firsthand collection of new data or new utilization of data already collected. In the former case it includes planning the field work and constructing schedules, questionnaires, or interview procedures so that the results obtained will throw light upon the specific problems or questions formulated in step 1. In the latter case it includes the location and selection of available data which can do the same. Here relevancy and design are the prime considerations. Mere collection of facts is not social research; the value and efficiency of the project are determined by the success with which the research worker is able to develop methods for getting information useful in answering his formulated questions. There are many ways for going about the getting of the required information. A discussion of several important types of collection of data will be found in the next chapter.

3. Collecting the data. In projects involving firsthand gathering of information the actual field work comprises this step. Such field work may be done either by the person conducting the research project or by other people. Certain principles and conventions have been developed, the ob-

servance of which makes for greater objectivity [4] in the results obtained. The most important of these principles is holding the getting at facts as the supreme goal; to achieve this, it is necessary for the scientific worker to discipline himself to put this goal before all other considerations— personal sympathies, allegiances to causes, and preferences for precon- ceived hypotheses. Absolute conformity to this principle is of course an impossible ideal, but the validity of any scientific research above personal impressions, popular opinion, or propaganda is based upon a relative ad- herence to it. Conventions of procedure have been developed which insure some basis for objectivity, although these have not reached the same de- gree of perfection in sociological research as in research in the physical sciences or even in psychology.

If already collected data are to be used, this step consists of actually taking them over from the published or unpublished records and also of investigating exactly how they were secured and of evaluating their re- liability. Expertness in evaluating data gathered by others can be best developed through experience in actual firsthand gathering of data during the period of training of the social research worker.

4. Classifying, editing, assembling, and tabulating data. The term "classifying" has several meanings according to time: the fundamental planning for classifying must be done in step 2, when the method is de- veloped by which the data from the field work are going to be used to throw light upon the question formulated; with this done, some classify- ing is often done in the actual field work, when units are enumerated as being in one class or another; finally, classifying in step 4 means the actual ordering of the data into classes by the processes of editing, assembling, and tabulating the data gathered. These last terms can be more easily defined than can classifying because of its several usages. Editing is the inspection of the records on which data have been gathered for mistakes, inconsistencies, and omissions. The results of editing a particular record may be approval, rejection, return to the enumerator for completion, or certain revisions by the editor, which have to be justified in any particular project. Assembling means the bringing together of the information con- tained on separate collection forms (registration blanks, schedules, or questionnaires) onto one sheet. In the modern punch card technique, as- sembling is dispensed with, or rather it is accomplished along with the process of tabulating. Tabulating means the actual listing of the data in classes.

5. Statistical analysis. When the data have been tabulated, they are ready for statistical analysis. While it is a commonly held opinion that

---

[4] We are using the term "objectivity" to mean a characteristic of results measured by the degree of agreement which there would be between these results and the results obtained by observation of the same phenomena by any other trained observer.

this is the merely routine part of a quantitative research project, the authors consider that here also there is afforded opportunity for the exercise of creative ability. Since the body of this text is occupied with expounding the methods of statistical analysis, we merely point out that this is the step where statistics can perform for sociology the functions outlined in Chapter 1. Here the summarizing measures are computed, or estimates are made of the summarizing measures for the universe; here relationships are investigated, and their summarizing coefficients determined or estimated, and hypotheses tested. In general, it is here that the mass of unintelligible data is analyzed and made to yield meaningful, condensed, and concise information. Again relevancy is extremely important as efficiency demands that only those analyses which bear upon the specific problem be computed.

6. **Presentation of results.**   The results of statistical analysis may be presented in textual form, either in words or in algebraic symbols; in tabular or semitabular form; and in graphic form, either by charts, diagrams, or modifications of standard graphic form, such as pictorial statistics or statistical maps. Sometimes rough drafts of presentation forms suggest interpretations to the investigator himself. Sometimes the presentation forms are not only of the results but also of the interpretations of them. At any rate, although much of the construction of presentation forms is purely mechanical, it is the extremely important step by which the research worker conveys to others whatever information he has gained by his inquiry. Excellence in this step demands varied abilities—clear and correct writing if the form is textual; careful organization if tabular; draftsmanship if graphic; and always judgment in selecting the best and most effective form of presentation for a particular set of results.

7. **Interpretation of the findings.**   This final step is of course the end to which all others lead. It should include a critical evaluation of the other steps of the project, a clear summary of the findings, and a translation of the statistical results into the terms of the original problem, thus giving whatever answers have been disclosed. This step calls for expertness both in statistics and in the field of application. When he has these qualifications, a research worker can confidently offer to the world his contribution to sociological knowledge by a careful interpretation of his findings.

Treatment of these seven steps in planning and executing a quantitative research project in this book.   The first step, the formulation of the research problem, is not treated in this book except insofar as it overlaps other steps. Content courses and general courses in social research are assumed to provide training in this matter. Steps 2, 3, and 4, the development of a method and the collection and classification of data will be treated briefly in the next two chapters. Step 5, statistical analysis, is of course the principal subject matter of all the book except Part I, and

so it is not treated in this part. Step 6 is treated in the last chapter of Part I. This inversion of logical order—treating the presentation of the results of statistical analysis before treating the actual methods of analysis —is made to introduce certain forms and conventions which should be utilized in all statistical work because they make for clarity in organization and for ease in conveying quantitative information to others. Finally, step 7, the interpretation of results, will be treated by illustrations throughout the book wherever examples of quantitative research projects are given since interpretation cannot be taught by a definite set of procedures or rules.

Flexibility of the pattern of social research. While we have divided the plan and execution of a quantitative research project into seven steps for the purpose of explanation, we do not intend to convey the idea that all research projects must conform to a stereotyped pattern. This idea is currently prevalent, especially among those who believe a "scientific" sociology will be achieved only by following the physical sciences closely in design of research projects.

It will be noted that the first step listed above was phrased, "formulating the specific questions of investigation in relation to the general problem or field of inquiry," rather than the more circumscribed, "formulating a hypothesis to be tested." The reason for doing this is that in the present state of sociology, adequate, scientific description of the phenomena in which sociology is interested should comprise a substantial portion of social research. Those who restrict the use of the term "research" to the testing of hypotheses, a phase of research which presumes a fairly advanced development of both theory and experimental techniques, are hardly being realistic about the present level of sociology or the indicated next steps for sociological research.

We repeat that the range of social research now needed is so great as to transcend any rigid outline of steps or any slavish imitation of procedures which may have proved fruitful in other fields. In certain projects adherence to such patterns may be useful, but the following of patterns should never be allowed to have a restrictive effect in discouraging experimental ventures in method. The more flexible social research can remain, the more chance it will have to utilize all possible contributions in procedures from other sciences and at the same time to invent more effective and appropriate procedures of its own.

SUGGESTED READINGS

Bernard, L. L., *The Fields and Methods of Sociology* (New York: Farrar and Rinehart, 1939).

Fry, C. Luther, *The Technique of Social Investigation* (New York: Harper, 1934).

Jahoda, Marie, Deutsch, Morton, and Cook, Stuart W., *Research Methods in Social Relations: With Especial Reference to Prejudice*, 2 vols. (New York: Dryden Press, 1951).

Lundberg, George A., *Social Research: A Study in Methods of Gathering Data* (New York: Longmans, 1942).

Myrdal, Gunnar, *An American Dilemma: The Negro Problem and Modern Democracy*, Appendix 2, "A Methodological Note on Facts and Valuations in Social Science" (New York: Harper, 1944).

Volkart, Edmund H. (ed.), *Social Behavior and Personality; Contributions of W. I. Thomas to Theory and Social Research* (New York: Social Science Research Council, 1951).

Young, Pauline V., *Scientific Social Surveys and Research: An Introduction to the Background, Content, Methods, and Analysis of Social Studies*, 2d ed. (New York: Prentice-Hall, 1950).

Sources and Collection of Data

I T W I L L be remembered that the first step in planning and executing a quantitative research project is the formulation of the specific question or questions to be answered or the setting up of a hypothesis to be tested. In research carried out during the period of graduate study, the services of the professor in the department who is most expert in the particular field of interest are usually available to the student for help in formulating his problem, as well as in supervising the actual research later on. Since this step is best advised by the one who knows the field of research most thoroughly and since its execution varies greatly with different fields, we are omitting discussion of it here and proceeding to the next step.

A comprehensive treatment of the sources of data for any one of the many special fields of sociology would constitute a book in itself. So, also, would a thorough treatment of the many and varied techniques and procedures used in the field collection of data. In this chapter we try only to give a basic outline of what these steps involve and to explain some of the terminology. Such an introductory treatment is necessary, however, to make clear the references to these phases of a quantitative research project in the more detailed treatment of the phase of statistical analysis. It is suggested that the student may well gain his initial familiarity with some of the most important sources of sociological data in getting his material for laboratory practice in applying statistical methods.

SOURCES OF DATA

Types of sources: primary and secondary. There are several classifications of sources which are useful in specifying the type of material used. First, there is the classification into primary and secondary sources. If the person or agency who has published data has either collected or supervised the collection of the data, the publication is called a *primary source*. Probably the most notable and valuable primary sources of soci-

ological data are the volumes of the decennial censuses of the United
States, containing data which are collected, edited, tabulated, partially
analyzed, and published by the Bureau of the Census of the Department
of Commerce of the United States government. If the person or agency
who has published data, or an analysis of data, has not collected or super-
vised the collection of the data, the publication is called a *secondary
source*. Examples of data in a secondary source are many of the tables in
All These People,[1] where figures by states, taken from the United States
census, have been grouped and combined into more meaningful regional
totals and averages. The feature which distinguishes these data as being
from a secondary rather than from a primary source is not their higher
degree of analysis and condensation, although this feature often char-
acterizes secondary source material, but is the fact that the author of
All These People did not collect the original data on which the tables re-
ferred to are based. Ordinarily, it is preferable to use primary rather than
secondary sources for obtaining data for further analysis since in primary
sources there is less chance of mistakes having crept in and since the data
are usually fuller and better adapted to further analysis. However,
secondary sources may be used if primary sources are no longer available
or if the secondary sources are in a more convenient form than the pri-
mary, so long as there is no question as to the scientific integrity of the
author. One will find gradations in the primary-secondary classification
which are intermediate and cannot be classified as strictly one or the other.

Types of sources: official and private.   Those data which are collected
by any governmental agency—federal, state, county, town, or municipal
—are called *official sources*. Those data which are collected by nongovern-
mental agencies or persons—businesses, private social agencies, non-
governmental research organizations, individual research workers—are
called *private sources*. Neither type is always more inherently trustworthy
than the other, since collection procedures vary among governmental
agencies as well as among private agencies and individuals. Because of the
expense and personnel required, only governmental agencies can ordi-
narily afford to make census type surveys over the entire nation, and thus
official sources often have the advantage of wider coverage. Again, gov-
ernmental agencies have the advantage of being able to require by law the
giving of information for certain censuses and the registration of certain
occurrences such as births, deaths, marriages, construction permits, etc.,
which registrations form the basis of certain official sources. Yet, certain
federal official statistics are reported on a purely voluntary basis, such as
parts of the Biennial Census of Manufactures, and certain local and even
state collections of data are poorly supervised and untrustworthy. No

---

[1] Vance, Rupert B., *All These People*: *The Nations Human Resources in the South* (Chapel
Hill: University of North Carolina Press, 1945).

classification label of a source can remove the necessity of examining how it was collected in order to evaluate its reliability.

Finding appropriate sources. With the increasing emphasis on records in both public and private administration, sources of data are becoming so numerous that it is very difficult to know when one has exhausted all the depositories of information on a particular subject. A few general suggestions can be offered. First, during his period of training for social research one should familiarize himself with the most important federal official primary sources. These include reports of the periodic censuses on population, housing, agriculture, manufactures, and business issued by the Bureau of the Census and also the reports based on the current population sample surveys of that agency; reports on vital statistics issued by the National Office of Vital Statistics; reports on matters pertaining to agriculture and the farm population issued by the Bureau of Agricultural Economics; reports on education, welfare, social security, and related subjects issued by the Federal Security Agency; reports on employment, wages, and cost of living of industrial workers issued by the Bureau of Labor Statistics; and reports on income and business conditions issued by the Department of Commerce. Some important sources of official data are listed at the end of this chapter. The student should become familiar with the various annual or biennial reports issued by the various departments or boards of his state on public health, welfare, and education and with the publications of his state planning board. With these most commonly used official sources in mind he will have some idea of what is available from governmental agencies when he begins a research project on a particular subject. Then he may begin a more exhaustive search through special official and private sources, using subject indexes in card catalogues of libraries and also author indexes if he knows the names of others who have done research in the field. Of course, in acquiring the background preparation for a research project the student should seek to obtain familiarity with the field, both as to sources and methods of analysis used by others, even though he intends to gather his own data firsthand.

COLLECTION OF DATA

When firsthand collection of data is indicated. When available sources do not supply the desired information, when the geographic area of study is sufficiently limited, and when financing of field work can be obtained, the research worker should plan to collect his own data. He should have explored thoroughly the existing sources to be sure that he is not about to perform a needless duplication. He should be sure that his time and money are carefully budgeted so that his project will not have to be stopped midway. Often projects of comprehensive scope can be broken

down into smaller units and executed one unit at a time; then if financial trouble arises, these units afford convenient stopping places and the work already done is not wasted.

In government projects or in those sponsored by large agencies the person planning and directing the research project may have field assistants who actually secure the information. However, since this text is primarily for beginning research workers who usually have to carry out themselves all phases of the research, their case is the one given primary consideration here.

Developing the method of obtaining facts to answer the formulated questions. In terms of time and dollars and results, it is always economical to plan thoroughly and carefully before field work is begun. Yet, since paper plans need to be checked against practical experience, it is necessary to intersperse some actual field work in the planning process. Developing a method should include the following specific matters of planning: selection of area, selection of sample, selection or construction of forms for collecting data, and modification of the forms in the light of preliminary field work. Since the selection of area is determined by the formulation of the specific problem, it will not be discussed here; since the principles underlying sampling can be understood only after probability and its use in statistics has been treated, they will be presented in Part III.

Selection and construction of collection forms. General texts on methods of social research usually have chapters devoted to the several types of collection forms, such as the schedule, the questionnaire, the registration blank, and the more flexible interview and observation techniques. We refer the reader to the following books for description of these forms and discussion of their respective advantages and disadvantages: Howard W. Odum and Katharine Jocher, *An Introduction to Social Research;*[2] Pauline V. Young, *Scientific Social Surveys and Research;*[3] George A. Lundberg, *Social Research;*[4] Mildred Parten, *Surveys, Polls, and Samples.*[5] We shall use here "collection form" as a generic term to include all types of forms and shall refer specifically to the "schedule," since it is the form most often used for collection of data for statistical analysis in sociological research. Often some type of collection form is used in combination with interviewing or observation techniques, especially where some of the material gathered is to be used for statistical analysis and some of it for nonstatistical study. At any rate, the type of

---

[2] *An Introduction to Social Research* (New York: Holt, 1929).
[3] *Scientific Social Surveys and Research: An Introduction to the Background, Content, Methods, and Analysis of Social Studies,* 2d ed. (New York: Prentice-Hall, 1950).
[4] *Social Research: A Study in Methods of Gathering Data,* 2d ed. (New York: Longmans, 1942).
[5] *Surveys, Polls, and Samples: Practical Procedures* (New York: Harper, 1950).

collection form selected—questionnaire, schedule, registration blank, or any combination or extension of these—should be chosen for its utility in getting the information required for the particular project.

If the preceding step in the research project has been adequately carried through—that is, the careful specification of the questions to be investigated and the exact definition of terms—the construction of the collection form will be greatly facilitated. We make certain suggestions for guidance in the construction of a collection form.

1. Examine all available collection forms which have been used in similar projects, noting especially their points of excellence for emulation and of inadequacy for avoidance.

2. Make the form as brief as possible and confine the questions to items relevant to the inquiry.

3. Avoid leading questions; avoid subjective rating scales or scales of any sort unless they have been standardized.

4. Write out an instruction sheet giving for each blank on the collection form the necessary definitions and specifications for all possible answers. Do this as carefully and in as much detail as if the enumeration were going to be done by untrained workers, even though the field work is to be carried out by the person constructing the form.

5. Wherever possible, have all possible answers indicated so that the enumerator can merely check the correct answer.

6. If the number of forms to be filled out is large (say 500 or more) and if arrangements can be made to secure the use of card punching, sorting, and tabulating machines, construct the schedules so that the cards can be punched directly from the collection forms with a minimum of coding.

7. Take psychological principles into consideration in arranging the order of the items on the collection form. Place first those which the interviewee would have no hesitancy in answering, and gradually lead up to those involving more personal information, such as income or sex life, which might end the interview if introduced too soon.

Pretesting and modifying collection forms. The advised procedure is to draw up roughly the forms for collecting data, intersperse a preliminary bit of field work for pretesting the forms, modify the forms in the light of the actual experience of the field work, try out again, modify again, and so on until a satisfactory form is developed. A rule which should be observed *invariably* is never to go to the expense of having forms printed until they have been given actual working tests in the field. The types of revisions which are often necessary after field tryouts are deletion of certain items on which it has been found to be impossible to get information, the change of classes to correspond with the actual classes of phenomena found, the addition of space for recording information which is offered and which is relevant to the inquiry, and the rewording of questions in the vernacular so they can be understood.

During the process of pretesting and modifying collection forms there should also be a testing of the procedure of assembling before the final form is adopted. Often in an initial attempt at assembling data gathered from preliminary field work, it will be discovered that the collection form on which the data are secured can be modified to make assembling much easier or even more accurate. Also the desired table forms should be drawn up in rough draft and a check made to see if the collection forms used are adequate for securing the results needed. This preliminary checking will not only make the field work go faster and more efficiently when once started, but it will also insure that the succeeding stages—assembling, tabulating, analyzing—will not be delayed or hampered by the lack of essential information or by the inclusion of unnecessary information. At this stage in some projects it is also advisable to do some preliminary statistical analysis. The particular case in which preliminary estimates of certain measures of dispersion are necessary to determine the size of the sample required will be treated in Part III.

Finally, the process of pretesting may make necessary some revision in the original formulation of the problem. Often even the questions investigated have to be changed when actual samples of data are available. While the steps of a research project were listed in the preceding chapter in a definite chronological order, at any one stage there must constantly be both a looking forward with preliminary planning of subsequent steps and a revision of previous steps in the light of the new experience. Probably the most critical point of a whole research project is at the stage of preliminary field work, where the good research worker by modifying all steps is able to plan finally the type and form of results he will be able to get as well as to restate his problem in terms of queries which are being answered by his data. The success of his whole project depends on his success here in achieving consistency and relevancy in the face of actual experience with the phenomena to be studied.

Preparation for collection of data. Certain things are assumed in the training of a person who is going to do field work; they can only be mentioned here. First is a familiarity with the subject matter under investigation. Unless the field work is to be a mere mechanical enumeration, the value of the results will be largely vitiated if the data are collected by one who does not understand them. Second, the principles and techniques of interviewing should at least have been read about in books dealing with such matters. It is to the advantage of one who is about to do research dealing with the responses of people to have studied these principles and techniques in social work or psychology courses. Unfortunately for students in sociology, the professionalization of social work training has meant the closing of many courses dealing with these aspects to all except those intending to enter the profession of social work. Third, honesty in

reporting is essential, although we offer no suggestions for its inculcation. Finally, there is required maturity and stability to a high degree. It is unfair to the people being studied to turn loose upon them research workers who may become emotionally upset and thus disturb the people or community functioning. While there is a valid place for social enlightenment, for promotion of social movements, and for agitation and effecting of social reforms, that place is definitely not in field work for social research. A high degree of self-discipline is required of the worker who is inclined to identify with the people he is studying and who, therefore, would like to change the distressing conditions he is investigating, but for the time must simply observe and record. Unless a prospective research worker can feel that the function of social research—finding the facts, analyzing, and interpreting them—has sufficient satisfactions in itself, he is probably not temperamentally suited to the task of social research. If his reaction is continually one of volubly deploring the futility of the routine of enumerating and calculating, while doing nothing about the social injustices he sees on all hands, he might better choose a vocation calling for more direct social action. But if he can take the long-time view that eventual action based upon scientific knowledge will be more effective and lasting, if he can exercise patience and restraint in the slow process of gaining such knowledge, both for itself and for its eventual use, then he has a right to feel that he is fulfilling as necessary a task in the general program of amelioration as those who actually accomplish the reforms.

The actual collection of data. No matter how carefully plans are laid, schedules are constructed, and work is organized, there will be during the period of field work a constant demand for the exercise of judgment, ingenuity, and adaptability in reorganizing procedures to fit the practical situation. The expert field person is guided by plans and rules; he observes carefully certain minimum requirements, but he never becomes a slave to formal procedures. If it were not for the magnitude of some projects, it would almost seem acceptable to say that the person planning and conducting the project should always do his own field work since in so doing he may observe and learn about his subject much more than just the quantitative facts enumerated on the schedules. The field worker should constantly observe and record all sorts of extraschedule material, which he may find extremely helpful in interpreting his quantitative results or in suggesting new hypotheses or new methods of measurement.

In spite of the necessity of flexibility of procedure, there are certain minimum requirements of good field practice which are so specific that they may be enumerated.

1. Never change the selection of units indicated when the sample is drawn. If unavoidable circumstances make impossible the study of certain indicated

units, form a plan for obtaining certain minimum control data about the non-interviewed units and for estimating their characteristics from data on interviewed units in the same classes according to the control data.

2. Always be honest with the interviewees. The type of problem and of people dealt with will determine whether or not a full statement and explanation of the research project can be made to the interviewees. But whether or not this can be done, whatever explanation is offered should not be deception.

3. The field worker should familiarize himself with the folkways and manners of the group he is studying and exercise these himself as far as is practicable. He should always have consideration for the time and work of the interviewees and should take care always to observe the customary courtesies.

4. Every blank on the collection form must be filled in. If necessary there can be differentiating symbols for such cases as, "interviewee refused to answer," "interviewee did not know the answer," or "interviewer did not ask for this information," but *never* should a space be left vacant. This rule is important in all forms of recording of social data.

5. Field work well done makes extraordinary demands on time, energy, and attention. The ideal situation is full-time field work with a strict personal regimen in order that the powers of observation and response may be kept at their highest degree of efficiency. Those suffering emotional disturbances, those in bad health, or even those who have not cultivated the ability of concentrating attention in spite of distraction are not in the condition demanded by social research field work.

SUGGESTED READINGS

Hauser, Phillip M. and Leonard, William R., *Government Statistics for Business Use* (New York: Wiley, 1946).

Payne, Stanley L., *The Art of Asking Questions* (Princeton: Princeton University Press, 1951).

Shryock, Henry S. and Lawrence, Norman, "The Current Status of State and Local Population Estimates in the Census Bureau," *Journal of the American Statistical Association*, 44 (June 1949), pp. 157–173.

SOURCES OF DATA

Federal Security Agency, National Office of Vital Statistics, *Summary of International Vital Statistics, 1937–1944* (Washington: Government Printing Office, 1947).

——, National Office of Vital Statistics, *Vital Statistics of the United States* (published annually).

——, Social Security Administration, *Social Security Bulletin* (published monthly).

Library of Congress, *National Censuses and Vital Statistics in Europe, 1918–1939: An Annotated Bibliography*, prepared by Henry J. Dubester (Washington: Government Printing Office, 1948).

United Nations, Secretariat, Statistical Office. *Demographic Yearbook* (published annually).

——, *Statistical Yearbook* (published annually).

——, *Monthly Bulletin of Statistics*.

United States Bureau of Agricultural Economics, *Agricultural Statistics* (published annually).

United States Bureau of the Census, *Catalog of United States Census Publications, 1790–1945* (Washington: Government Printing Office, 1950).

——, *Catalog and Subject Guide of United States Census Publications* (Washington: Government Printing Office, published quarterly).

——, *Current Population Reports* (Series on various subjects; some are issued monthly, some annually, and others at irregular intervals).

——, *Statistical Abstract of the United States* (Washington: Government Printing Office, published annually).

——, *Historical Statistics of the United States, 1789–1945: A Supplement to the Statistical Abstract of the United States* (Washington: Government Printing Office, 1949).

——, *County Data Book: A Supplement to the Statistical Abstract of the United States* (Washington: Government Printing Office, 1947).

——, *Cities Supplement to the Statistical Abstract of the United States* (Washington: Government Printing Office, 1944).

United States Bureau of Labor Statistics. *Monthly Labor Review.*

United States Department of Commerce. *Survey of Current Business* (published monthly).

Assembly and Tabulation of Data

THE three processes—editing, assembling, and tabulating data—constitute the intermediate step between actual field collection of data and the statistical analysis of them. Through these processes the more inclusive process of classification is accomplished. The processes vary so much with the nature and size of the research project that a short chapter cannot begin to cover all contingencies of actual practice. We shall define the processes and describe the procedures they involve for relatively simple cases. Throughout the chapter we deal primarily with the techniques used when punched-card technique is not available, since it is likely that the simpler type of assembling and tabulating will be the first with which the student will have experience.

EDITING

The process of editing. When all the steps of a research project are done by the same person, this step is not of great importance. It consists simply in checking over all the entries on the collection form to make sure that all blanks are filled in and that no absurd or highly improbable entries have been made. In projects utilizing many enumerators, especially untrained ones, editing is very important. The editor examines each schedule for completeness, consistency, and plausibility. Then he makes a decision that the schedule is to be rejected, returned to the enumerator for changes or completion, accepted as it stands, or accepted with certain revisions which the editor makes himself. Some major surveys have been severely criticized for "overediting," that is, for changing too freely the figures entered by the enumerator to "bring them into line." The difficult question of the degree of editing required and permitted is tied up with the matter of using untrained enumerators and is a persistent problem in connection with all research projects where the enumeration is done by comparatively untrained workers. We only point

out and do not dwell upon this problem because we are dealing principally with the case where one person does all the work.

An example of the editing process is the fact that in the 1940 and 1950 census reports there are no persons listed with age unknown. Whenever a schedule is returned with age unknown or not given, an editor fills in an estimate of the age. This estimate is based on the other information about the person, such as age of spouse, relation to head of household, years of school completed, etc.

Coding. An extension of editing is "coding," which is required in punched-card technique and is sometimes used if assembling is to be done by hand sorting. Coding means the assigning of numbers to classes in order that the number may be punched onto cards or may be used in other ways to designate each class. For instance, in the schedule referred to below calling for information on race, the two classes "white" and "nonwhite" might be assigned the numbers "1" and "2," respectively. In such a case the number is actually written on the schedule for each item. Editors' and coders' marks are usually made in pencil or ink of a different color from that used in the original filling in of the schedule.

ASSEMBLING

Assembling, as has already been explained, is the process of transferring the data gathered from the individual collection forms to one or more assembly sheets. The most common practice is to use large assembly sheets with vertical and horizontal lines drawn or printed on them. It is possible to get pads of prepared sheets, called analysis pads, or columnar pads, which have as many as 45 rows and up to 40 columns. One or more columns is assigned each item of information called for on the schedule. The information from any one schedule is then entered in one row, a horizontal line being allowed for each schedule.

For a nonquantitative characteristic, such as color, we would assign a column to each possible response, such as white and nonwhite, and place a check mark in the appropriate column for each schedule. For a quantitative characteristic, such as age, we have two alternatives.[1] We can set up class intervals, assign a column to each class interval and merely check the appropriate column for each schedule, or we can use only one column and write in the actual figure.[2] Conditions to be fulfilled in setting up class intervals are discussed in Chapter 8.

---

[1] For a fuller definition and illustration of the types of characteristics sociologists study see the classification of characteristics on pp. 66–67.

[2] We shall use the term "category" to refer to the classes of a nonquantitative classification, the term "class interval" to refer to the classes of a quantitative classification, and the term "class" to refer to either type or both types.

TABULATING

Tabulating data relating to nonquantitative characteristics. Tabulating is an initial process in summarizing all of the data from individuals on any single item.[3] It differs according to the type of assembling which has been done and according to the nature of the item being tabulated. The simplest case of tabulation is the tabulation of data relating to one nonquantitative characteristic. This is done by counting the tally entries in the column corresponding to each category.

Cross-tabulation of data relating to two or more nonquantitative characteristics. Cross-tabulation means a joint tabulation of the data relating to two or more items of information. Suppose in our tabulation we have the item "sex" with two columns, "male" and "female," and also the item "color" with "white" and "nonwhite." If we wish to cross-tabulate these two items, we must count the individuals that are checked in both the "male" and the "white" columns in order to get the number of "white males." We must repeat this procedure for the other combinations in order to get the number in each sex color group. If we wish to cross-tabulate by another item, such as "residence," at the same time in order to get the number of "rural white males," etc. the process quickly becomes very laborious and quite difficult to perform accurately. Two procedures can make the process shorter. The first is a different method of assembling whereby all the combinations such as "rural white male," "rural white female," etc., are written at the top of columns instead of (or in addition to) the single category column headings, and the individual schedules are tallied for combinations instead of (or as well as) for single categories. Then tabulation is accomplished by simply counting the tallies in the columns with the combination headings. Thus, before assembling is begun, the necessary cross-tabulations should be decided upon in order that the most efficient procedure may be determined in advance.

Hand and machine sorting. A second procedure may be used where the schedules are sturdy enough to stand handling or where their data are recopied onto cards. Sorting these schedules or cards now takes the place of assembling, but sorting must be repeated for each item. In the above illustration, the cards or schedules are first sorted into two piles, "white" and "nonwhite"; next, each of these piles is sorted into two piles, "male" and "female"; and finally, each of these four piles is sorted into two piles, "rural" and "urban." Then tabulation consists of counting and listing the number of cards in each of the eight combinations. This

---

[3] Individual, as used in this chapter, means the individual unit investigated for which one schedule is filled in. The individual may be a family, a school, a farm, a community, or any other social unit, as well as a person.

method of assembling and tabulating may be done by hand, but where there are many schedules, it is also quite laborious. Machines have been developed for carrying through the process in three steps. First, the data of the coded schedules are punched onto cards; next, these cards are run through a sorting machine; next, the sorted piles are run through a machine which counts and tabulates them. The more recently developed machines are highly automatic and combine the last two steps. For cross-tabulation each pile is again sorted for a second item unless machines are used which perform cross-tabulations at one operation.

Another device used in hand sorting is a margin-punched card. These cards have a row of holes near the edge and a separate category is assigned to each hole. For example, holes 1, 2, and 3 might stand for urban, rural-nonfarm, and rural-farm, respectively. If the schedule being transferred to the card is marked *urban*, hole number 1 would be punched out so that it becomes a U-shaped gap in the edge of the card instead of a hole near the edge. When the cards are all punched and stacked an ice pick or knitting needle device can be run through hole number 1 and the cards lifted. The cards punched for *urban* will fall and the others will remain on the needle. This supplies a rapid device for sorting and cross-tabulating. Margin-punched cards are adapted to either nonquantitative or quantitative data and also furnish a visual means for detecting relationships between variables.[4]

Tabulating data relating to quantitative characteristics. As explained in the paragraph on assembling, the tabulation of data relating to quantitative characteristics is facilitated if the assembling has been done by making tallies in columns designated for each class interval of the variable. If in the case of the age of workers such a procedure of assembling has been carried out, tabulation is accomplished by counting the tallies in each column and listing their totals in a column with each entry on the same horizontal line as its class interval limits, which are placed in the leftmost column. Such a listing of class interval limits and the frequencies of the classes (the numbers of individuals) in two parallel columns is called a frequency table. Frequency tables are not only a convenient method of actually presenting findings, but are also the most common way in which data relating to a quantitative characteristic are prepared for further statistical analysis. Fuller treatment of the form of frequency tables will be given in Chapter 5 and of their construction, meaning, and use in Chapter 8.

Sometimes, however, it is not desirable to use the procedure of tallying

[4] *Sixteenth Census of the United States, 1940.* Agriculture. Handbook, Uses of Agriculture Census Statistics (Washington: Government Printing Office, 1943), pp. 225–245; *Keysort Punching and Sorting Manual* (New York: The McBee Company, 295 Madison Avenue, 1950).

in class interval columns when assembling. It may be that we want to have the exact value of the variable written on the individual's line for some later treatment, such as ungrouped correlation analysis. In such a case only the actual numbers are entered on the assembly sheet, and for tabulation into a frequency table there must be an extra step of tallying these numbers onto another sheet, called a tally sheet, which has class intervals specified and from which the totals can be arranged into a frequency table as described above.

Cross-tabulation of data relating to two quantitative characteristics. Cross-tabulation of data on two quantitative variables is almost always done on a separate sheet rather than on the original assembly sheet. In cross-tabulation the class intervals of one quantitative variable form the column headings, and the class intervals of the other variable form the leftmost entries on the horizontal lines. For each individual a tally mark is made in the cell which is under the appropriate column heading and on the horizontal line designated with the appropriate interval. A two-way frequency table resulting from the tallying of a cross-tabulation has approximately the same form as the tally sheets with numbers (totals of the tallies) in each cell rather than the tally marks. Examples of such two-way tables can be found in the chapter on correlation—in fact, such a two-way table may be called a correlation table. Cross-tabulation can also be done by sorting the schedules or cards, first according to the class intervals of one variable and then sorting each pile for the class intervals of the other variable.

Cross-tabulating data relating to quantitative and nonquantitative characteristics. It may be necessary to cross-tabulate the data relating to a quantitative characteristic, such as age, with those relating to a non-quantitative characteristic, such as sex. Again, we have a choice of procedures: the listing of ages (or tallying by age groups) can be done for each sex separately when assembling the data, it can be done afterwards on a separate tally sheet, or it can be done by sorting the cards or schedules twice. It is the equivalent of making as many tabulations of the quantitative variable as there are categories in the nonquantitative classification. It is highly important to think through in advance just which cross-tabulations are going to be needed so that the whole assembling and tabulating processes can be planned most efficiently.

SUGGESTED READINGS

Armstrong, Lawrence W., "The Application of Punch Card Equipment to Statistical Processing" (Washington: Agricultural Research Administration, 1951, *mimeographed*).

Hartkemeier, Harry P., *Principles of Punch-card Machine Operation* (New York: Crowell, 1942).

Keysort Punching and Sorting Manual (New York: The McBee Company, 295 Madison Avenue, 1950).

Pease, Katharine, *Machine Computation of Elementary Statistics with Special Reference to the Friden, Marchant, and Monroe Calculating Machines* (New York: Chartwell House, 1949).

Presentation of Results

I F A L L the steps so far described have been carried out for a quantitative research project, the next step is the statistical analysis of the tabulated data. All the succeeding chapters of this book will be devoted to the methods of statistical analysis. In this chapter, however, we are going to imagine that the steps of statistical analysis and of interpretation have been completed and that the person executing the project is now ready to arrange his findings into the form by which he can present his results to others. The availability and usefulness of whatever contribution to scientific knowledge the research project has yielded are dependent upon the successful execution of this final step in the project.

Conventions in forms of presentation. In any one of the special forms of presentation of quantitative material the chief desiderata are logical organization, preciseness, and ease of comprehension. In order to achieve these most efficiently—and along with them other desirable qualities, such as economy of space, clarity, and correct emphasis—certain conventions in structure and style of the form of presentation have been developed. Many of these conventions are based upon the principles of logical and systematic arrangement; many of them are purely arbitrary. Some of them are almost universally observed; others are less rigorously followed. On points where there is option, research organizations or administrative departments often set up arbitrary rules to secure uniformity.

The general nature of the report of a quantitative research project. The choice of presentation forms, the degree of fullness and detail, the style of writing, and the general make-up of a report will vary according to the nature and purpose of the research project. The findings may be typed and submitted as one unit to the director of a cooperative research project to be incorporated by him later, along with other units, in a publication. Or the findings may be typed in the particular form required by a university for a thesis or a dissertation. They may be typed or mimeographed and submitted to a committee or some other body which

has commissioned the research. Or they may be typed and submitted directly for publication as an article, monograph, or book. Almost always, however, the research worker types or supervises the typing of his findings before he hands them over to someone else, whether it be to a thesis committee, an organization, or a publisher. Therefore, our chief emphasis will be on the conventions observed in preparing a typescript of the findings of a statistical project. The general rules for manuscript preparation apply, of course, but the inclusion of quantitative material introduces many special problems not ordinarily covered in general manuals of form, most of which are designed for printed material.

The major forms of presentation. The findings of a quantitative research project may be presented in *textual*, *tabular*, or *graphic* form, or in modifications or combinations of any of these. Certain parts of the report of any quantitative research project are almost invariably presented in textual form—the statement of the problem, the description of the method of study, and the interpretation of the results. If supplementary nonstatistical research findings are included, these, too, are usually given in textual form. Such parts are governed by the ordinary rules of composition, which we shall not go into except to emphasize the necessity for precise, unambiguous, and simple composition. A mastery of the fundamentals of grammar and punctuation is, of course, prerequisite to achieving clear writing.

We are concerned more especially with those parts of the report which offer quantitative data and the results of their statistical analysis. First, there is the matter of choice of form for presenting any set of data or of statistical results—textual, tabular, or graphic, or a combination of any two of these or of the three. No determinate rules can be laid down for making the choice of presentation form, but the following general principles should be borne in mind when one is choosing.

1. Tables are the most concise and efficient form for presenting quantitative data which are numerous or detailed.

2. Unless the important features of the data in tabular form are obvious, some textual matter pointing out these features should supplement tables.

3. Graphic presentation is primarily for the purpose of calling attention to the grosser features of the findings. Ordinarily charts and maps should be accompanied by tables supplying the same information in more detail.

PRESENTATION OF RESULTS IN TEXTUAL FORM

Conventions regarding numerals. When detailed data are not desired in the presentation, the results of a statistical project may be given entirely in textual form. In such cases the question of usage most frequently arising is when to spell out numbers and when to use numerals.

Unless an organization or agency for which the report is being made has adopted its own rules of usage, the chapter on "Numerals" in the *Style Manual* issued by the United States Government Printing Office [1] is advised as a guide. The general principle is that numbers smaller than 10 are to be spelled out, while those as large as 10 are to be written as numerals, thus,

Of the three incorporated places in Orange County, Chapel Hill with a population of 9,169 is the largest.

There are many exceptions to this general principle. Numbers of 10 or more at the beginning of a sentence are usually spelled out unless they contain a decimal. Numbers smaller than 10 are written as numerals when they refer to serial numbers, time, or dates, when they have decimals in them, or when they are used in a series in combination with numbers of 10 or more. For example,

Table 6 shows the age distribution of all third-grade children by two-year intervals for January 1, 1940, as well as their mean age of 9.3 years and their range in age from 6 to 14.

Conventions regarding algebraic and other symbols. If summarizing measures or estimates of them have been determined, it is permissible to include these in textual material with only the value of the measure given in numerical form or to write out an equation using algebraic symbols and numerals, thus,

The coefficient of correlation between the age of the tenant farm women and the number of grades of school they have completed is —.49.

Or,

$r_{XY} = $ —.49, where X is the age of the tenant farm woman and Y is the number of grades she has completed.

Regression equations are best written with the algebraic symbols, usually on a line to themselves and indented, thus,

$$X_1 = 1.043 - 1.279X_2 - 2.065X_3$$

A letter used as an algebraic symbol should be underlined in a typescript so that it will appear italicized when printed, as in the equations above.

When equations or formulas appearing in the text have equal marks, square root signs, subscripts, Greek letters, and other symbols not appearing on the keyboard of an ordinary typewriter, the writer has a choice of three procedures: using a typewriter which has a keyboard equipped with algebraic and statistical symbols; using a typewriter with a standard keyboard for the parts of the equation or formula it will write

---

[1] *Style Manual*, rev. ed. (Washington: Government Printing Office, 1945).

and filling in the other parts by hand; or writing all the algebraic or symbolic parts by hand.

Comments on tables. When text is used to explain, comment upon, or emphasize tables, the textual material should not merely duplicate the data in the tables, but it should point out relations, changes, and other important features. For instance, the following sentences might well accompany Table 1 illustrated later in this chapter:

The total population of the United States increased nearly fivefold between 1860 and 1950, but the Negro population in the United States increased only a little over threefold in the same period. The number of foreign-born whites has not increased as rapidly as the total population.

This sort of textual supplementation is by far preferable to the stereotyped sort which simply duplicates the information in the table.

Semitabular or leader form. A form intermediate between textual and tabular is called semitabular or leader form. It is treated here because it is more nearly like textual and is less complex than tabular. This form is used for material too brief to be put into a formal table but too long to be included in the regular text. Page 312 includes an example of semitabular form.

It is not advisable to use semitabular form a great deal because it does not specify nearly so completely as regular tabular form the identification of the data as to time and source. Only if such specifications are clearly stated in the accompanying text is semitabular form permissible in research reporting.

PRESENTATION OF RESULTS IN TABULAR FORM

Statistical tables. An orderly arrangement of numerical data into rows and columns, with concise labels specifying the nature of the data, is called a statistical table, or simply a table. Tables are the most generally useful and the most indispensable form of presentation of the quantitative results of a research project. Good table construction requires careful planning for each individual table as to the best method of classification and arrangement for making its presentation clear, logical, and concise. Learning the conventions pertaining to table making is an essential requirement of training in the use of statistics in social research. For a treatment of these conventions and their alternate forms and the principles of logic involved in classification, the reader is referred to *Statistical Tables: Their Structure and Use*,[2] by Helen M. Walker and

---

[2] *Statistical Tables: Their Structure and Use* (New York: Teachers College, Columbia University, 1936).

Walter N. Durost, *Bureau of the Census Manual of Tabular Presentation*,[3] and *Handbook of Tabular Presentation*,[4] by R. O. Hall.

Types of Tables. Tables are classified as "general-purpose" tables or as "special-purpose" tables according to their function. General-purpose tables ordinarily contain quite detailed information and are intended for use as reference. Special-purpose tables ordinarily give a condensation or an analysis of data and are therefore sometimes called "summary" tables. The line of demarcation between the two is not hard and fast, and it is not necessary to emphasize the differentiation here since the conventions of table form apply equally to both.

Parts of a table. Tables are composed of at least four parts: heading, caption (or box), stub, and body. Notes may be considered a fifth part although they are not always present. The *heading* consists of the serial number of the table, the title of the table, and sometimes a prefatory note. All tables in any sort of report should be numbered consecutively throughout the typescript in the order in which they appear. Roman numerals were formerly used for these serial numbers, but the less cumbersome Arabic numerals are now preferred. The title of a table should be carefully worded to state the necessary specifications of the data as concisely as possible, but without abbreviations or symbols. For population material and for many other sorts of sociological subject matter, it is often convenient to follow the sequence of answering the questions, "What?" "Where?" and "When?" in the title of a table. The *caption* of a table consists of the headings over each column of data. The *stub* of a table consists of the designations of the horizontal rows of data, and it is placed on the leftmost edge of the table. A table usually involves two classifications, the classes of one classification appearing in the caption and those of the other in the stub. Certain types of classifications are more commonly found in one or the other place, but usually convenience in spacing requires that the classification with the greater number of classes appear in the stub. The *body* of the table consists of the actual figures ordered into appropriate rows and columns. Ordinarily only figures or symbols appear in the body of the table. *Notes* may appear in various places. Notes pertaining to the entire table may appear as headnotes between the title and caption. Explanatory footnotes may either appear in the body of the table just above the ruled lines with the source notes just under the rules, or the footnotes may be placed below the ruled lines immediately preceding the source notes.

---

[3] *Bureau of the Census Manual of Tabular Presentation* (Washington: Government Printing Office, 1949).

[4] *Handbook of Tabular Presentation: How to Design and Edit Statistical Tables* (New York: Ronald, 1943).

Table 1. POPULATION OF CONTINENTAL UNITED STATES BY RACE
WITH NATIVITY FOR WHITES, 1860–1950

[Figures given to nearest 1,000 population]

		White				
Year	Total, all races	Total	Native	Foreign-born	Negro	Other races
1950a	150,697	135,215	125,068	10,147	14,894	588
1940	131,669	118,215	106,796	11,419	12,866	589
1930	122,775	110,287	96,303	13,983	11,891	597
1920	105,711	94,821	81,108	13,713	10,463	426
1910	91,972	81,732	68,386	13,346	9,828	412
1900	75,994	66,809	56,595	10,214	8,834	351
1890	62,948	55,101	45,979	9,122	7,489	358
1880	50,156	43,403	36,843	6,560	6,581	172
1870	38,558	33,589	28,096	5,494	4,880	89
1860	31,443	26,922	22,826	4,097	4,442	79

a. Preliminary figures.
Source: Historical Statistics of the United States,
1789–1945 (Washington: Government Printing Office, 1949),
p. 25; 1950 Census of Population, Preliminary Reports.
Series PC-7, No. 1. "General Characteristics of the Population of the United States: April 1, 1950," Table 1.

Table 2. POPULATION OF CONTINENTAL UNITED STATES BY RACE WITH
NATIVITY FOR WHITES, 1860–1950

[Figures given to nearest 1,000 population]

		White				
Year	Total, all races	Total	Native	Foreign-born	Negro	Other races
1950 a	150,697	135,215	125,068	10,147	14,894	588
1940	131,669	118,215	106,796	11,419	12,866	589
1930	122,775	110,287	96,303	13,983	11,891	597
1920	105,711	94,821	81,108	13,713	10,463	426
1910	91,972	81,732	68,386	13,346	9,828	412
1900	75,994	66,809	56,595	10,214	8,834	351
1890	62,948	55,101	45,979	9,122	7,489	358
1880	50,156	43,403	36,843	6,560	6,581	172
1870	38,558	33,589	28,096	5,494	4,880	89
1860	31,443	26,922	22,826	4,097	4,442	79

a Preliminary figures.
Source: Historical Statistics of the United States, 1789–1945 (Washington: Government
Printing Office, 1949), p. 25; 1950 Census of Population, Preliminary Reports. Series PC-7,
No. 1. "General Characteristics of the Population of the United States: April 1, 1950," Table 1.

Typing of tables. Since the research worker is usually more concerned with getting his tables into proper typed form than with the technical details of printing, we have used a typed table in accepted form as Table 1. Table 2 is the same table in printed form. Other examples of printed tables can be seen throughout the text.

To avoid the many problems which arise when text and tables are interspersed on the same page, it is best to type each table on a separate

Table 3. FREQUENCY DISTRIBUTION OF THE 443 ECONOMIC AREAS[a] OF THE UNITED STATES BY CLASS INTERVALS OF PERCENTAGE CHANGE IN POPULATION, 1940–1950

Percentage change in population, 1940–1950	Number of economic areas
All percentages	443
−30.0 – −20.1	2
−20.0 – −10.1	19
−10.0 – − 0.1	81
0.0 – 9.9	121
10.0 – 19.9	100
20.0 – 29.9	47
30.0 – 39.9	23
40.0 – 49.9	17
50.0 – 59.9	19
60.0 – 69.9	7
70.0 – 79.9	3
80.0 – 89.9	2
90.0 and over [b]	2

[a] See Donald J. Bogue, *State Economic Areas: A description of the procedure used in making a functional grouping of the counties in the United States* (Washington: U. S. Government Printing Office, 1951).

[b] The mean value of these two cases is 118.1.

Source: 1950 Census of Population. Preliminary Counts. Series PC-3, No. 7, "Population of State Economic Areas: April 1, 1950."

page and to insert it immediately following the page where the table is first mentioned. Insofar as possible the top and side margins of pages on which tables are typed should be the same as those of pages containing text. For tables containing very few columns, such as Table 3, the margins are usually made wider in order to bring the columns closer together and to make a neater appearing table. A widely followed convention is always to have the table appear vertically on the page to avoid the inconvenience of having to turn a volume sideways to read the table. If necessary, the

table may be typed on a sheet of extra width and folded in. This is also an advisable procedure in order to avoid having a table continued on a second page. Every table should have a complete source note unless all of the data have been collected firsthand by the writer. To provide a complete reference and at the same time to provide for the greatest possible brevity in source notes, agencies or organizations may set up certain rules for content and form of source notes.

Other suggestions about constructing tables. In the process of preparation of a table one or more dummy forms should be constructed to try out the most effective arrangement of the data and to determine the space requirements. It is well to have someone entirely unfamiliar with the findings being presented look over the proposed table to check on its clarity. Although the accompanying text may contain a more detailed explanation of the steps of analysis used in arriving at the results presented in a table, a table should be complete in itself. Every table should be able to stand alone and to convey its information without any ambiguity or possibility of incorrect interpretation.

PRESENTATION OF RESULTS IN GRAPHIC FORM

Advantages of graphic presentation. It is in terms of numbers that quantitative data are presented in textual or tabular form, and the comprehending of data so presented requires a mental process of imagining numbers of units or magnitudes of characteristics corresponding to the numbers in the text or table. This process becomes difficult when one tries to imagine more than a few numbers or quantities simultaneously for the purpose of comparison, especially since other differentiating characteristics must be held in mind at the same time to distinguish the quantities being compared.

In graphic presentation of quantitative data, magnitudes and quantities are represented not by abstract numbers, but by geometric designs, where the length or area of a part of a presentation form is proportional to the quantity or magnitude represented and where position, color, design, or other differentiating characteristics of the presentation form represent quantitative or nonquantitative differences in the phenomena being described. Thus, instead of the bare numbers and words or tables and text, visually observable differences in quantity and symbolic representation of other differences facilitate the comprehending of the data. The mental process of reconstructing the quantitative aspects of the actual sociological situation from which the data were gathered is aided by careful graphic presentation of the results of statistical analysis of those data. Such presentation supplies other cues for reconstruction than those contained in abstract numbers. It is obvious that the less experi-

enced the reader is with mathematics and the study of numerical relations, the greater the advantages of graphic over tabular presentation for him. This advantage is really twofold, for it means that the reader grasps certain comparisons and relations which he would never get from table or text and also that he grasps quickly other comparisons and relations which he would get only from long study of the other forms of presentation.

Since the mere format of graphic presentation differentiates it from textual material, it draws attention, and since good graphic presentation conveys information almost at a glance, another of its advantages is that the results so presented are much more likely to be noted by others. In fact, an attractive interspersing of textual with graphic material may inveigle hesitant readers into reading a whole report.

Limitations of graphic presentation. The first limitation of graphic presentation is that it is valuable chiefly for conveying information about relative quantities, that is, for making comparisons. Without accompanying numerals graphic presentation cannot show absolute magnitudes. To present graphically the information that the United States in 1950 had a population of 150,697,361 would be difficult and pointless. However, a graphical presentation with two bars whose heights are proportional to the population of the United States in 1940 and in 1950 probably conveys a better idea of the comparison between the two than do the mere numbers, 131,669,275 and 150,697,361.

In addition to being limited mainly to conveying relative information, graphical methods are limited in the number of facts that can be presented at one time. Tabular form should be used if it is necessary to present a great many detailed facts. From such tables careful selection should be made of the important features and relationships, and graphic presentation should be used to emphasize these relationships which might otherwise be lost in the mass of detail.

Types of graphic presentation. While the types of graphic presentation are rather clearly differentiated, there is no close agreement among writers as to the names to be applied to the different types. The *Statistical Dictionary of Terms and Symbols*[5] by Albert K. Kurtz and Harold A. Edgerton lists chart, diagram, figure, and graph as all synonymous, although noting that chart is sometimes used in contradistinction to graph. The choice of names used here is somewhat arbitrary, but no type names of graphic forms seem to be universally used.

The first type of graphic form we shall consider is the *bar diagram*, where the length of the bar is proportional to the quantity represented. Secondly, we shall consider *area diagrams*, where the area of the figure is

---

[5] *Statistical Dictionary of Terms and Symbols* (New York: Wiley, 1939).

Figure 1. Native White, Foreign-born White, and Nonwhite Population in the United States, 1950. (Source: Table 1.)

Figure 2. Native White, Foreign-born White, and Nonwhite Population in the United States, 1860–1950. (Source: Table 1.)

proportional to the quantity represented. Thirdly, we shall consider *coordinate charts*, often called *graphs*, where distances from two coordinate axes are proportional to two quantities. Fourthly, we shall consider *statistical maps*, which show geographic distribution of characteristics. And finally, we shall consider briefly *pictorial statistics*.

Bar diagrams.   Simple bar diagrams are adapted to present cross-classified data when one of the classifications is quantitative and the other is either nonquantitative or is the class interval groupings of a quantitative variable. The lengths of the bars are proportional to the quantitative characteristic, while the different bars represent the categories or class intervals of the other characteristic. Figures 1, 2 and 3 illustrate several degrees of complexity of bar diagrams. Notice that in Figure 1 the bars are horizontal, as is customary when the different bars represent categories of a nonquantitative classification. Vertical bars are used in Figures 2 and 3 since each bar represents a class interval of a quantitative characteristic.

The width of the bars and the space between the bars has no statistical meaning. The spaces between the bars should not be the same width as the bars, because if the bars are unshaded, spaces may be confused with bars. Vertical lettering is always to be avoided so that the reader will not have to turn the diagram sidewise to understand it.

Figure 3. Percent of United States Population Native White, Foreign-born White, and Nonwhite, 1860–1950. (Source: Table 1.)

Population pyramids are an example of bilateral bar diagrams where the bars are contiguous. Figure 4 presents population pyramids for the United States in 1940 and 1950 superimposed. Table 4 shows the data necessary in order to construct these population pyramids. The student should note that in computing the percentages necessary for constructing a population pyramid, the total population, combined male and female,

Table 4. DISTRIBUTION OF UNITED STATES POPULATION BY AGE AND SEX, 1940 AND 1950
[Figures given to nearest 1,000]

Age	1940			1950 [a]		
	Total	Male	Female	Total	Male	Female
All ages	131,669	66,061	65,608	150,697	74,633	76,064
0–4	10,542	5,355	5,187	16,324	8,301	8,023
5–9	10,685	5,419	5,266	13,241	6,825	6,416
10–14	11,746	5,952	5,794	11,361	5,680	5,681
15–19	12,333	6,180	6,153	10,732	5,302	5,431
20–24	11,588	5,692	5,896	11,327	5,457	5,870
25–29	11,097	5,451	5,646	12,093	5,924	6,169
30–34	10,242	5,070	5,172	11,601	5,735	5,866
35–39	9,545	4,745	4,800	11,193	5,476	5,717
40–44	8,788	4,419	4,369	10,058	5,029	5,029
45–49	8,255	4,209	4,046	8,990	4,520	4,470
50–54	7,257	3,753	3,504	8,274	4,036	4,238
55–59	5,844	3,011	2,833	7,230	3,608	3,622
60–64	4,728	2,398	2,330	5,950	3,029	2,921
65–69	3,807	1,896	1,911	5,060	2,364	2,696
70–74	2,570	1,271	1,299	3,425	1,610	1,815
75 and over	2,643	1,239	1,404	3,837	1,737	2,100

[a] Preliminary figures.

Source: *1950 Census of Population, Preliminary Reports.* Series PC-7, No. 1. "General Characteristics of the Population of the United States: April 1, 1950," Table 1.

is the 100-percent base. If each sex group were made to total 100 percent, then sex differences in the total population would not show in the population pyramid.

Area diagrams. A common form of area diagram is the pie diagram (sometimes called a pie chart or a sector chart), which serves the same function as a percentage-composition bar diagram, such as Figure 3. Figure 5 shows some of the same information that is shown in Figure 3. This is quite an effective way of presenting composition for one year, but when a series of such circles are made to correspond to the bars of Figure 3, it is less effective in showing composition than the series of bars since

Figure 4.   Age and Sex Distribution of United States Population, 1940 and 1950. (Source: Table 4.)

the areas to be compared are not so placed as to facilitate contrast. It is possible to convey the information of Figure 2 with a series of pie diagrams whose radii are proportional to the square root of the population. Figure 6 is of this type. It again is more difficult to interpret quickly than the corresponding bar diagram.

Coordinate charts. The term "coordinate" is chosen to designate this type of chart in view of its origin in analytic geometry, where points are located with reference to two perpendicular axes called coordinates, and lines or curves are drawn to connect such located points. Such charts are often called line charts, curve charts, or simply graphs. Whatever term is used to designate them, these charts are useful for presenting data on the cross-classification of two quantitative variables. The two major sub-

Figure 5. Proportions of United States Population Classed as Native White, Foreign-born White, and Nonwhite, 1860, 1900, and 1950. (Source: Table 1.)

types are the one where for any value of the first variable, there may be any number of values of the second variable, and the other where for any value of the first variable (often called the "independent" variable), there is only one value of the second variable. The first type is usually known as a correlation chart, or a scatter plot or diagram, and illustrations of this type will be deferred until the chapter on correlation. Common examples of the second type are charts of time series and charts of frequency distributions.

Figure 6. Population of the United States with Composition by Color and Nativity, 1860, 1900, and 1950. (Source: Table 1.)

A time series is the record of a sequence of observations made at specified intervals of time. The data of Table 5 are a time series in which the time intervals are five years and the quantities observed are the number of immigrants to the United States from Italy and Ireland in one year. Figure 7 shows a coordinate chart presenting the information for Ireland graphically. It will be noted that there are two scales, the horizontal

Figure 7.   Immigrants to the United States from Ireland, by Five-year Periods, 1870–1945. (Source: Table 5.)

representing time and the vertical representing number of immigrants. Time is conventionally expressed on the horizontal scale, which is usually reserved for the "independent" variable. For any one year a point is plotted directly over the point representing the midpoint of that year on the time scale and at a distance above the scale equivalent to the distance on the vertical scale representing the number of immigrants during that year. The points thus located are connected with straight lines. The graphing of a time series is important in connection with the statistical analysis of the fluctuations of a time series into their components.

It is often desirable to compare two time series for the same period of time and for the same type of phenomena but for different areas. Figure 8 is a coordinate chart displaying the two time series of Table 5, immigrants from Ireland and from Italy. The pattern of immigration from Ireland is the same as in Figure 7 except that the vertical scale has been com-

Figure 8.   Immigrants to the United States from Ireland and Italy, 1870–1945. Arithmetic vertical scale. (Source: Table 5.)

Table 5. IMMIGRANTS TO THE UNITED STATES
FROM ITALY AND IRELAND, 1870–1945

Year	Country	
	Ireland	Italy
1945	427	213
1940	839	5,302
1935	454	6,566
1930	23,445	22,327
1925	26,650	6,203
1920	9,591	95,145
1915	14,185	49,688
1910	29,855	215,535
1905	52,945	221,479
1900	35,730	100,135
1895	46,304	35,427
1890	53,024	52,003
1885	51,795	13,642
1880	71,603	12,354
1875	37,957	3,631
1870	56,996	2,891

Source: *Historical Statistics of the United States, 1789–1945* (Washington: Government Printing Office, 1949), pp. 33–34.

pressed in order that the larger number of immigrants from Italy might be plotted on the same coordinate chart. Figure 8 provides a graphic comparison of the number of immigrants from the two countries.

In analysis of time series data, however, interest usually focuses upon the changes, and Figure 8 may suggest erroneous interpretations of the rates of change in number of immigrants from the two countries. For example, the increase in number of immigrants between 1875 and 1880 appears considerably greater for Ireland than for Italy, as indeed it is. But if the two percentages of increase are computed from Table 5, it can be seen that the greater *amount* of increase in immigrants from Ireland represents only an 89 *percent* increase, while the smaller *amount* of increase in immigrants from Italy represents a 240 *percent* increase. Therefore, a type of chart called the semilogarithmic [6] has been developed which makes possible a direct comparison of percentages of change or rates of change.

In the semilogarithmic chart the horizontal scale and distances re-

---

[6] Graph paper which has a regular scale in one direction and a logarithmic scale in the other is called *semilogarithmic*, while graph paper which has logarithmic scales in both directions is called *double logarithmic*. The terminology is, of course, inconsistent.

main as they are in the chart just described, but the vertical distances are plotted proportional to the logarithms of the numbers to be represented rather than proportional to the absolute numbers. If one were to look up the logarithms of the number of immigrants for each of the years for which observations are recorded in Table 5 and to plot distances proportional to these above the appropriate year, he would have a semilogarithmic chart of the time series under discussion. A much simpler way of accomplishing the same thing is by the use of semilogarithmic paper, which has the lines in one direction equally spaced but in the other direction spaced so that the distance from the scale reading to the reference line is proportional to the logarithm of that scale reading. Figure 9 shows the results obtained by plotting the data of Figure 8 on a set of axes with a logarithmic vertical scale. The only new procedure in making a semilogarithmic chart is that of selecting and specifying the vertical scale. In the present illustration this is done as follows: First, the lowest and highest values to be shown on the vertical scale are found from Table 5 to be 213 (Italy, 1945) and 221,479 (Italy, 1905). The larger of these values is over 1,000 times as great as the smaller, and it is from this relation that we select the number of cycles we shall need on the logarithmic scale. Several types of semilogarithmic paper are available, differing in the number of cycles they contain. Each cycle consists of 10 horizontal lines numbered 1, 2, 3, 4, 5, 6, 7, 8, 9, 1, and the last line of one cycle is the first line of the next cycle. One-cycle paper will show data where the largest value is not more than 10 times as great as the smallest value; two-cycle paper will show data where the largest value is not more than 100 times as great as the smallest value; three-cycle paper will show data where the largest value is not more than 1,000 times as great as the smallest value; and so on, with each cycle covering a range of values 10 times as great as that covered by the preceding cycle. Since our largest value is more than 1,000 times as great as the smallest, we see that we cannot represent our data on either one-, two-, or three-cycle paper, although we shall not need all of the fourth cycle.

Next, we must assign a value (greater than zero) to the first line of the first cycle, sometimes called the reference line. The value is usually taken to be some multiple of 10, and if convenient, a power of 10. The value assigned must be lower than the lowest value to be plotted; so, in our case we have chosen 100. We could have used 200, but the plotting is much easier if the value chosen for the reference line is a power of 10. After the first line of the lowest cycle has been assigned a scale value, the scale values for the other lines of that cycle are determined by multiplying the scale value of the reference line by the successive numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. In Figure 9 this determines values from 100 to 1,000 for the lines of the first cycle. Beginning anew with the second cycle, the

Number of Immigrants

Figure 9. Immigrants to the United States from Ireland and Italy, 1870–1945. Logarithmic vertical scale. (Source: Table 5.)

scale value of its first line, 1,000, is multiplied by successive integers from 1 through 10, determining values from 1,000 to 10,000 for the second cycle. Similarly, the third cycle has values from 10,000 to 100,000 and the fourth cycle has values from 100,000 to 1,000,000, though all of the fourth cycle is not shown.

Next, we simply plot points for each country for each year using the scale values just determined to measure their vertical distance. In Fig-

UNITED STATES TOTAL
5,859,169

I DOT = 500 FARMS
(COUNTY UNIT BASIS)

Map 1. Number of Farms, Jan. 1, 1945. (Source: Bureau of the Census.)

ure 9 equal proportions or percentages of change are represented by equal vertical distances, and, therefore, the percent increase from 1875 to 1880 is shown as greater for Italy than for Ireland.

Because equal vertical distances on semilogarithmic charts always represent equal percentages of change, semilogarithmic charts are to be used where the rate or proportion of change is the feature to be emphasized. Because equal vertical distances or ordinary charts (called "arithmetic" charts in contradistinction to semilogarithmic charts) represent equal amounts of change, arithmetic charts are to be used when the amount of change is the feature to be emphasized. A special advantage of the semilogarithmic chart is its usefulness in showing comparisons in rates of growth or change for two different time series on the same chart. Even when the two series vary greatly in absolute amounts of quantities, they can be shown on the same semilogarithmic chart since two different scales may be used on a logarithmic vertical axis.

Another important way to present the data of a frequency distribution is with a histogram. The histogram is actually an area diagram rather than a coordinate chart, but since it is used to represent frequency distributions, we mention it here. A series of rectangles are drawn with their bases on the horizontal axis. A rectangle is drawn for each class interval with its width equal to the class interval and its area proportional to the frequency of the class. If the class intervals are equal, the height of each

RANK OF COUNTIES

☐ Lowest fifth
☐ Next lowest
▨ Middle fifth
▨ Next to highest
■ Highest fifth

Map 2. Farm Operator Families' Levels of Living, 1945. (Source: Bureau of Agricultural Economics.)

rectangle is also proportional to the frequency of its class. Figures 10 and 12 of Chapter 8 are histograms.

Statistical maps. When results to be presented consist of quantitative data for geographic units, statistical maps are often an effective mode of presentation. Because maps portray certain geographic aspects such as distance and continuity, statistical mapping of the distribution of a characteristic may show patterns, gradients, and other spatial relations which a table cannot disclose. There are three major types of statistical maps: in the first type a dot or a spot representing some stated number or quantity is placed on a map to show the location of the phenomena; in the second type the geographic units are grouped into classes according to the class intervals of the measure of the characteristic, and each geographic unit on the map is hatched or colored with the design or color assigned to its interval; in the third type some sort of chart or diagram

portraying data relating to a specific geographic unit is superimposed upon that unit.

The first type of map presents difficulties in the spacing of the dots. Usually it is assumed that the number of dots to be placed on one geographic area can be spread evenly over it without doing violence to the facts being represented. Of course, the smaller that the geographic area is, the more nearly this assumption holds. Map 1 is a dot map of farms in the United States, January 1, 1945, on a county unit basis.

The second type of map is more generally used. The number of intervals into which the geographic units are grouped is usually from two to eight, and there are two ways of making the grouping. The units may be put into order from the one having the highest measure to the one having the lowest measure and then divided into four, five, or more groups of equal numbers. If there are four such groups, they are called quartiles, if five, quintiles, etc. Or the units may be grouped according to the class intervals in which their measures fall, when the class intervals are set up arbitrarily, usually equal. Maps showing quartile groupings have the advantage of showing finer differentiation in the part of the range where the concentration is since no more than one fourth of the units will be shaded similarly; but they have the disadvantage of giving no information about where these points of concentration are unless they are accompanied either by a table or by a legend which would necessarily be awkward because of the irregularity of the intervals. A method of grouping with equal-size class intervals has the advantage of showing at what ranges the measures are concentrated, but it has the disadvantage of not differentiating so finely as the quartile method the variations of states within the range of concentration. The disadvantage of the quartile method is largely overcome if a table accompanies the map; the disadvantage of the equal interval method can be somewhat overcome by increasing the number of intervals until the differentiation required is achieved. Map 2 is made by the quintile method and shows farm operator families' levels of living in 1945.

Map 3 illustrates the third type of statistical map where diagrams or charts are superimposed upon the geographic areas or locations to which they refer. It is a particularly appropriate type for contrasting distributions of characteristics in different regions or locations, which is often the essential point of interest in studies in human geography or human ecology. The actual making of such a map involves chiefly the techniques of diagram or chart construction, since these are simply pasted onto a base map in their appropriate places.

Pictorial statistics. A modification of the horizontal bar diagram has been developed chiefly by Otto Neurath and by Rudolf Modley. The horizontal bars are replaced by rows of symbols picturing the phenomena

Map 3. Percent Increase in Urban and Rural Population of the United States by Regions, 1940 to 1950. (Source: *1950 Census of Population, Advance Reports.* Series PC-8, Table 1.)

whose quantity is being represented. Rules and conventions for constructing pictorial diagrams may be found clearly expounded in Modley's *How to Use Pictorial Statistics*,[7] while a slightly different set of principles is explained and illustrated in Neurath's *International Picture Language, The First Rules of Isotype.*[8]

Practical suggestions for the construction of graphic presentation forms. After a selection has been made of the data to be put into graphic form, and the type of graphic form appropriate to the material has been chosen, the actual laying out and drafting of the presentation form is the next step. As in the case of field work, it is an important phase in training for social research to learn how to plan and construct graphic forms. Even though one may be able to afford to hire an expert draftsman to construct his forms, he will always be able to plan better the type and details of his graphic presentation if he has learned to construct such forms himself.

Graphic forms of presentation, unlike tables, are not reset by a printer when a report is handed in for publication, but they are reproduced without change except for enlargement or reduction. Therefore, a writer of a report has the responsibility of preparing his graphic forms exactly as they are to appear when published. Since when a person prepares a report, he often does not know whether it will be kept on file indefinitely in typed form or will be published, it seems the wisest course to prepare graphic forms which will be appropriate for either disposal. The following suggestions are offered for the preparation of forms which will be acceptable either in bound typescripts or for submitting to a publisher.

Materials and equipment.[9] For diagrams and charts use smooth drawing paper, 20- or 24-pound smooth white bond, or engineer's tracing paper if the chart is to be traced; or use a coordinate paper with non-photostatic blue lines if the chart is not to be traced. For maps, if possible, choose or have made base maps which are 8½ by 11 inches or which are of such dimensions that they will easily fold in this size in a bound typescript.[10]

All lettering or lines which are to be reproduced must be drawn with India ink. Straight lines require a ruling pen; curved lines and letters are best made with a lettering pen. The Leroy lettering set is recommended for making all letters and numerals, since its use can be mastered quickly by one who is not a draftsman. The Leroy sets are quite expensive, however, and there are other cheaper lettering sets which can be used if one

---

[7] *How to Use Pictorial Statistics* (New York: Harper, 1937).

[8] *International Picture Language, The First Rules of Isotype* (London: George Routledge, 1936).

[9] All equipment and all materials mentioned (except Zip-A-Tone) can be purchased from Keuffel and Esser, Hoboken, New Jersey.

[10] A wide variety of base maps and of coordinate paper is available from Codex Book Co., Inc., 74 Broadway, Norwood, Mass.

takes the time to acquire skill in their use. Freehand lettering is difficult to learn, and if one has not already learned it and does not have a lettering set available, a typewriter is the best alternative. If a typewriter is used to make the letters and numerals on a chart, a fresh black ribbon should be used and a carbon sheet should be put behind the chart with the carbon side next to the chart to make the letters and numerals appear blacker and reproduce better. While good lettering done either with a lettering set or by an expert in freehand lettering is the first preference, neatly typed lettering is preferable to irregular and inexpert hand lettering.

For the hatching on maps and on certain diagrams the waxed sheets of designs manufactured by the Paratone Company [11] are recommended. These sheets, or screens as they are called, are applied to a drawing, the part of the drawing which is to be hatched is traced around with a cutting pen, the surplus is removed, and the applique burnished down with a bone instrument. The process can be learned much more easily than the process of hatching with a ruling pen and India ink, and the expense of the waxed sheets is not great when the time saving is taken into account. If these prepared hatchings are not available, the beginner should choose very simple designs for hatching at first.

Conventions relating to the layout of a graphic presentation form. As with tables, the top of all graphic forms should be at the top of the page to prevent the reader's having to turn the volume sidewise to look at them. Likewise, every word or numeral on the figure should be placed in a normal upright position. Unlike tables, however, charts and diagrams (except pictorial charts) often have their titles underneath the figure rather than above it. Maps have no set place for titles or legends; some authors prefer placing the title of a map at the top, some prefer the bottom, and others place the title at the top, bottom, or side of the map—wherever the irregularities of the map leave the most room. Sources of data should *always* be indicated for graphic forms; although if there is an accompanying table, the reference may be simply to this table. Source notes are usually written in lower-case letters, immediately following the title, and are usually enclosed in parentheses.

Every graphic form should be completely self-explanatory. If a form is so complex that it requires elaborate textual explanation, it cannot serve the primary function of graphic presentation—to convey information quickly and easily.

In charts and diagrams scales should be clearly indicated. The designation of the nature and units of the vertical scale should be lettered just

---

[11] Most of the maps and shaded diagrams in this text were made with Zip-A-Tone. The sheets may be purchased from Paratone Company, Inc., 440 South Dearborn Street, Chicago, Illinois.

over the left border with the leftmost letter lined up with the leftmost numeral of the vertical scale and with the first letter of the serial number preceding the title. The designation of the nature and the units of the horizontal scale is centered, usually below a coordinate chart and below or above a horizontal bar diagram. Careful thought must be given to the selection of title and scale designations for each chart, because sometimes the title carries part of the specification of the nature of the scale clearly and sometimes it does not. Such an infinite variety of types of characteristics and of their units of measurement are possible, that no rules can be given except that title, vertical and horizontal scale designations, and other legend material should be planned so that together they clearly specify exactly what the figure represents. Since the individual words in scale designations (which are analogous to captions in tables) are not to be emphasized, they are lettered in lower-case letters with only the first word capitalized.

If several lines or curves are shown on one coordinate chart, they should be differentiated by making them of combinations of straight lines and dots. Crosses or circles should not be used for this purpose as they generally indicate particular points of interest. It should be remembered that for differentiating either lines or areas on charts or maps, colors should not be used if the form is to be reproduced, unless one intends to have the much more expensive type of color reproduction.

Finally, it is unwise to attempt to present in graphic form too much information or too exact information. For detailed presentation tables are needed. A graphic form containing too many lines or curves or too many numerals is confusing and does not meet the chief criteria of good graphic presentation—selection of important features for emphasis, clarity, and ready comprehension by the reader.

SUGGESTED READINGS

Brinton, Willard Cope, *Graphic Presentation* (New York: Brinton Associates, 1939).

Bureau of the Census Manual of Tabular Presentation (Washington: Government Printing Office, 1949).

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chaps. III, IV, V, and VI.

Hall, R. O., *Handbook of Tabular Presentation* (New York: Ronald, 1943).

Lutz, R. R., *Graphic Presentation Simplified* (New York: Funk and Wagnalls, 1949).

The Preparation of Statistical Tables, A Handbook (Washington: Bureau of Agricultural Economics, 1949).

Style Manual, rev. ed. (Washington: Government Printing Office, 1945).

Weld, Walter E., *How to Chart; Facts from Figures with Graphs* (Norwood, Mass.: Codex Book, 1947).

PART II
Descriptive Statistics

~~~~~~
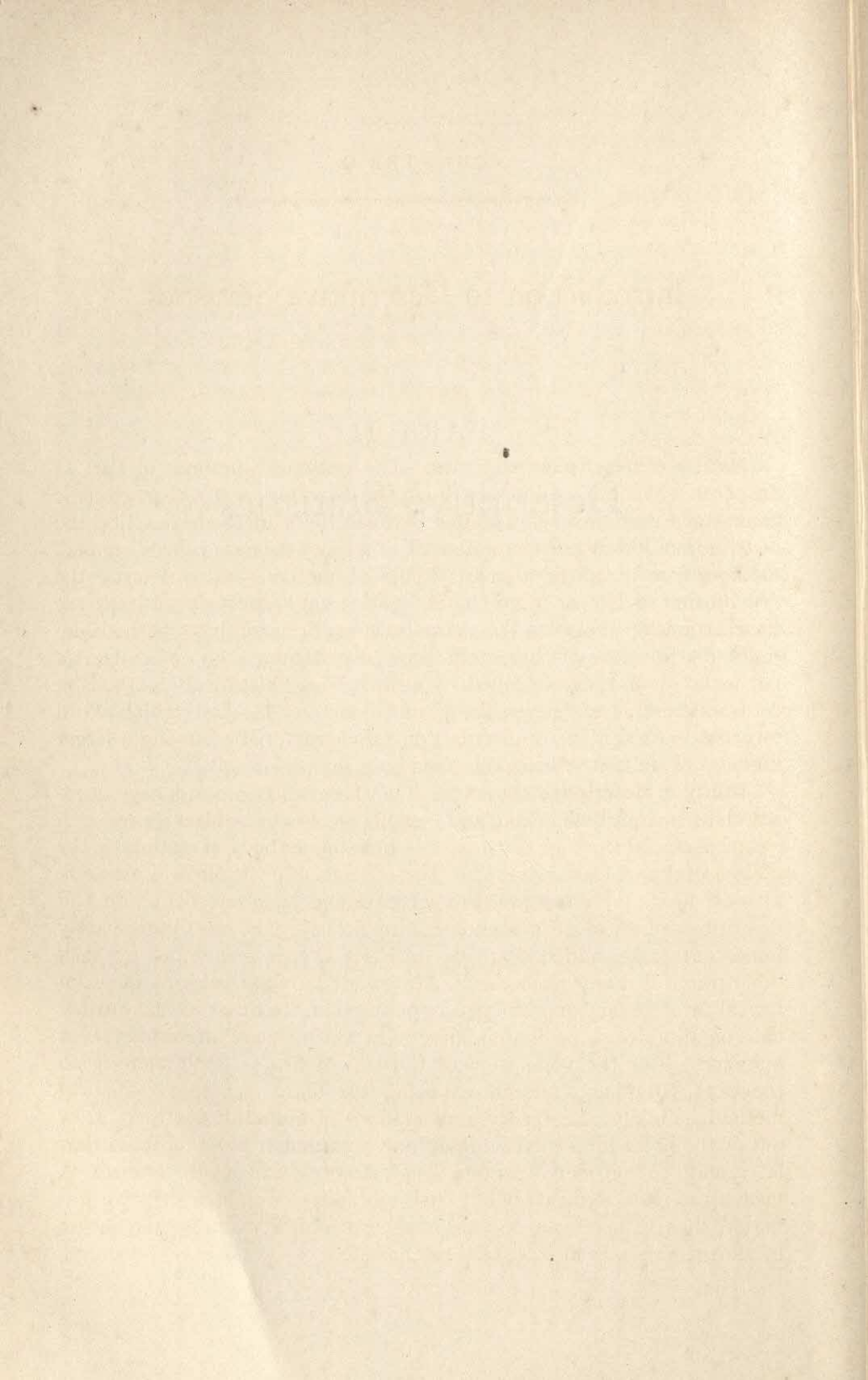
Introduction to Descriptive Statistics

Nature of descriptive statistics. The statistical methods of Part II are those which are used in describing the distribution of a characteristic among a series of varying units. Because these methods describe the distribution of only one characteristic at a time, they are called "simple" methods in contradistinction to "complex" methods which describe the distribution of two or more characteristics simultaneously and analyze the relationships between the characteristics. Because these methods describe distributions of characteristics of a particular group of units at a particular time, they are called "descriptive" or "historical" methods in contradistinction to "generalizing" or "inductive" methods which afford estimated descriptions of distributions of characteristics among a larger universe of units from data observed on a sample of units.

Utility of descriptive statistics. For those who do not engage in research as such but who make and compile records to be used for research —such as social work administrators—this simpler body of methods is the most useful part of statistics. For those engaged in sociological research, at least in its present state, the accurate and adequate definition and description of sociological phenomena are probably the most immediately important tasks, and descriptive statistics affords techniques for such description of many phenomena. Moreover, in many subjects of sociological interest, and particularly in population, we often have available data on all units of the limited universe in which we are interested (as for instance, all of the population of the United States), and there is no practical sampling situation requiring the more elaborate statistical methods. Finally, the theory and methods of inductive statistics grow out of the procedures used in describing a particular group of units that later may be considered as the sample from which one generalizes. A mastery of these elementary but basic methods of simple descriptive statistics, then, is necessary because of their own utility and is prerequisite to all further study in statistical methods.

Conditions for using descriptive statistics. In problems of statistical description those properties, traits, or attributes which "characterize" an individual unit may be called "characteristics." As long as interest in a characteristic is limited to whether or not a single unit possesses it or to how much of it a single unit possesses, the problem of description is not statistical. When, however, there is interest in the way in which a characteristic is distributed among a number of units, the problem becomes statistical. Again, if perfect uniformity with respect to a characteristic prevails throughout the units of a group, the description of the distribution of the characteristic in that group does not require statistical methods. But if some units possess the characteristic and others do not, or if some possess it in greater degrees than others, that is, if the units of a group vary with respect to the characteristic, a second criterion for the use of statistical methods is fulfilled. Thus, we see that *plurality* of units investigated and their *variation* in characteristics are two of the basic conditions for the applicability of statistical methods to a research problem. Since sociologists deal primarily with *groups* which manifest *variation* on at least two levels—the variation among units within a group, and the variation among groups—most of their research problems meet these two basic conditions of the applicability of statistical methods. The third basic condition, however, which is *enumerability* or *measurability* of the characteristics studied, is not met by certain of the problems which are of interest to sociologists. Whether this is due to inherent nonmeasurability of the subject matter or simply to the undeveloped stage of sociology in perfecting measuring instruments is not agreed upon by all sociologists. At any rate, for the description of groups in regard to enumerable or measurable characteristics, with respect to which the units of a group show variation, descriptive statistics affords the most concise and adequate methods available.

Object of statistical description. Let us note that while the unit possessing a characteristic is the observable object, it is the similarities and differences of units with respect to various characteristics which are the object of statistical description, and, indeed, of scientific inquiry generally. The object of statistical description is not the observed unit but is more abstract—the distributions of characteristics among the units. It is true that complete descriptions of the distributions of all characteristics of a group of units would also be a complete description of the group of units. But for such complex units as those with which sociologists deal—persons, families, farms, neighborhoods, institutions, states, etc.— *complete* descriptions of *all* characteristics is a Utopian concept. Short of a stage of such perfection, the available descriptions of distributions of characteristics among units are not complete descriptions of the units, al-

though they may be complete descriptions of the distributions of certain characteristics. In the process of statistical description, then, the emphasis is on the *characteristics* rather than on the *units* which manifest them. All scientific thought or study requires this same process of abstraction from the concrete manifesting unit to the abstract manifested characteristic. In interpretations of statistical descriptions, however, we shall often speak of "descriptions of groups of units," thus translating our abstract descriptions of distributions into more concrete terms. Research reporting is pedantic if it does not attempt to relate its findings to everyday concepts and meanings, even though the process of statistical analysis itself is necessarily performed at an abstract level.

Types of characteristics. Since the distributions of characteristics are the object of statistical description, the type of statistical methods to be selected for describing the distribution of a particular characteristic depends upon the type of the characteristic. In order to be able to specify the type of characteristic for which the various statistical methods are appropriate, we shall expand the tentative definitions given in Chapter 4 into a more complete classification of characteristics by type. We are not concerned here with the controversial question of whether all phenomena of sociological interest are potentially amenable to statistical description. We are attempting to classify only those characteristics which are actually or potentially measurable or enumerable.

The basis for differentiating types of characteristics is the nature of their measures of incidence for a single unit—the form of the data obtained by observation and other techniques. The classification made on this basis is shown on pages 66 and 67.

Other classifiers call type I characteristics "qualitative" as contrasted with the quantitative characteristics of types II and III. This designation has the advantage of suggesting the idea of differences in kind for qualitative characteristics and of differences in degree or amount for quantitative characteristics. We shall use the term when we wish to convey this suggestion. Unfortunately, however, the word "qualitative" is also frequently used to mean a type of characteristic which is not amenable to statistical description, and it is to avoid any such suggestion that we choose instead to use "nonquantitative," although this choice might be subject to the same criticism.

The differentiation between the characteristics of type I and those of types II and III is due to inherent differences in the nature of the two types of characteristics, while the differentiation between the characteristics of type II and those of type III may be due to inherent differences, or it may be due to differences in the state of scientific advance in the study of the two types of characteristics. A large amount of effort is being

CHARACTERISTICS CLASSIFIED BY THE

Type of characteristic

I. NONQUANTITATIVE CHARACTERISTICS

 A. Dichotomous (sometimes called all-or-none variables)

  Example: Sex—Male or not-male

 B. Manifold classifications

  Example: Regional location—Northeast, Southeast, Southwest, Middle States, Northwest, Far West

II. QUANTITATIVE CHARACTERISTICS FOR WHICH PRECISE MEASURING DEVICES HAVE NOT BEEN DEVELOPED

 A. Those for which only rough classes of degree of incidence can be differentiated
  Example: Condition of housing—Good, fair, poor

 B. Those for which relative degree of incidence can be differentiated as more than or less than another degree, affording ranks for series of units
  Example: Eminence of a group of scientists

 C. Those for which measuring scales have been developed, but with intervals which have not been demonstrated to be equal
  Example: Attitude of unfavorableness toward the Negro

III. QUANTITATIVE CHARACTERISTICS FOR WHICH MEASURING DEVICES PROVIDE MEASURES WITH UNITS EQUAL AND ADDITIVE

 A. Those for which incidence is measured in integers

  Example: Fertility (number of children ever borne)

 B. Those for which finely graduated degrees of incidence can be measured (measures made on a theoretically continuous scale)

  Example: Age

NATURE OF THEIR MEASURES OF INCIDENCE

Nature of measure of incidence of characteristic for a single unit	Nature of measures of incidence of characteristic for a series of units (Data on the distribution of the characteristic among a plurality of units)
Enumeration of presence or absence of major category	Two integral numbers denoting frequencies in the two categories
Indication by check mark of either "male" or "not-male"	In United States 1950: 74,633,000 males; 76,064,000 females
Enumeration of presence of one of several mutually exclusive traits	Integral number denoting frequency for each category
Indication by check mark of the incidence of the trait of being located in the Northeast, Southeast, etc.	In United States 1945: 6,977,000 passenger cars in Northeast; 3,790,000 Southeast, etc.
Designation of class within which degree of characteristic falls	Integral number denoting frequency for each category
Indication by check mark of condition of house as good, fair, or poor	In township 4: houses with condition good, 121; fair, 382; poor, 369
No absolute measure possible for a single unit	Array of units by rank in characteristic, giving a series of ordinal numbers
	Scientist X, rank 1; Scientist Y, rank 2; etc.
A quasi-cardinal number representing a scale value	A series of quasi-cardinal numbers representing scale value for each unit (these may be grouped into a frequency distribution)
A quasi-cardinal number representing score on an attitude test	Individual X, scores 2.3; Individual Y, score 7.9; etc.
A cardinal number, an integer	A series of integral cardinal numbers representing scale value for each unit (these may be grouped into a frequency distribution)
A cardinal number representing number of children ever borne	Woman X, 4 children; Woman Y, 2 children; etc.
A cardinal number which may take any value within the limits of precision of the fineness of measuring device chosen	A series of cardinal numbers which may take values differing from one another by as little as the "least count" of the measuring device (these may be grouped into a frequency distribution)
A cardinal number representing years lived, which may be expressed to any chosen number of decimal places.	Individual A, 41.8 years; Individual B, 31.4 years; etc.

directed toward the development of more precise measuring devices. Success of these efforts will enable us to classify as of type III characteristics which are now regarded as of type II.

Differentiation between nonquantitative and quantitative characteristics. The finer differentiation of the subtypes of types II and III are not so important to the student now as they will be when he is studying statistics of relationship. The essential differentiation which he should understand at this stage is that between nonquantitative and both types of quantitative characteristics. We do not mean to imply that there is a hard and fast line of demarcation since, on the one hand, determination of qualitative differences is often by quantitative methods and, on the other hand, quantitative differences or differences in degree may be so great as to be in effect qualitative differences or differences in kind. Ideally, however, by a nonquantitative characteristic we mean an attribute which a unit either possesses or does not possess; and by a quantitative characteristic we mean a property which presumably all units possess, but in varying degrees, which are potentially or actually measurable. A nonquantitative characteristic is studied by enumeration, or count, of the individuals who have a certain attribute; the second type of characteristic is studied by measuring the degree of the characteristic possessed by each individual. Data from a study of a nonquantitative characteristic are assembled and tabulated into nonquantitative categories; data from a study of a quantitative characteristic are assembled and tabulated into class intervals of a frequency distribution or into some other of the various forms used to describe the distribution of a quantitative variable. The incidence among a series of units of a nonquantitative characteristic, about which data are secured by enumerating the units of the series which possess the characteristic, as well as the incidence of a quantitative characteristic, about which data are secured by measuring the amount of the characteristic possessed by each unit in the series, is called the "distribution" of the characteristic.

Illustration of differentiation between types of characteristics. As an illustration of a nonquantitative characteristic, we may consider the attribute of "blindness." In studying the incidence of this attribute in a group, we first inspect all individuals of the group and make a record of whether each is "blind" or "not-blind." In assembling the data we have two categories for this item and label one column "blind" and another "not-blind," and make a tally mark for each individual in the proper column. Tabulating these data consists of counting all the tally marks in each column and the process results in two figures, one the number of blind individuals and the other the number of not-blind individuals. These figures are then ready for further analysis or condensation such as the determination of ratios, or proportions, or percentages.

As an illustration of a quantitative characteristic, we may consider the property of "age." In studying the distribution of this quantitative variable in a group, we determine the age (probably to the last birthday) of each individual and record the age on the collection forms. In assembling and tabulating we follow either of the two procedures explained in the chapters on assembling and tabulating and arrive at a frequency distribution of all the ages or at a list of the ages, either of which is ready to be analyzed or condensed further.

The difference between quantitative and nonquantitative characteristics is not so clear-cut, however, as the above incomplete illustration suggests. Let us notice the quantitative considerations involved in the illustration about the incidence of blindness. First, there is the matter of definition of blindness. Shall we count only those individuals who have absolutely no sight as "blind?" Or shall we include also those who can distinguish light from dark but who cannot distinguish objects? Or shall we include also those who can distinguish gross objects but who cannot see well enough to read even newspaper headlines? Actually, the working definition of "blindness" usually employed is the possession of not more than one tenth of normal sight, as measured by a standardized procedure.[1] Thus we see that a "qualitative" attribute may be "quantitatively" defined, since in practice we must measure the individual as to sight before we can enumerate him as "blind" or "not-blind." Let us suppose further that blindness rates have been determined for demographic units and that we wish to study them comparatively. We may now think of "blindness ratio" (number of blind individuals per 1,000 population) as a quantitative characteristic for which the different demographic units have varying values. When we change the order of varying unit, we change the type of the characteristic being described from nonquantitative to quantitative. Thus we see that there is no absolute dichotomy into qualitative and quantitative characteristics, although it is convenient to consider them as of a different nature and to treat them differently.

On the other hand, in the illustration regarding age we may have made the study for the purpose of seeing what proportions of the population fall into the three groups: children, workers, and old people. We may define children as those of age 0–17, workers as those of age 18–64, and old people as those over 64. If we tabulate the quantitative data into these three classes, we may afterwards prefer to think of the classes as categories which are qualitatively different since the individuals in each category have different functions. It is the same as if we had originally enumerated all the individuals as possessing or not possessing the nonquantitative characteristics of being a child, or a worker, or an old person.

---

[1] Harry Best, "Blindness: Definition and Statistics," *American Sociological Review*, 4 (August 1939), pp. 488–492

Descriptive methods for different types of characteristics. Thus, quality and quantity are seen to merge into one another and not to be necessarily opposing concepts. Yet, the two types of characteristics are usually distinguishable and are more conveniently studied by the methods appropriate for one or the other type, even though there may be a change from viewing a characteristic as one type to viewing it as another type in the same project. Certain statistical methods apply to the description of the distribution of nonquantitative characteristics, certain apply to the description of the distribution of quantitative characteristics, and certain are applicable to both. The various methods by which summary descriptions of groups of units as regards their nonquantitative and quantitative characteristics are made comprise the subject matter of Part II.

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
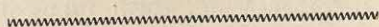
Nonquantitative Distributions: Ratios, Proportions, Percentages, and Rates

THE methods of this chapter are primarily, although not exclusively, adapted to the description of the distributions of nonquantitative characteristics. First, ratios, proportions, and percentages are explained and illustrated as summarizing measures which condense information on the incidence of a nonquantitative characteristic among a group of units. Following the presentation of this primary use, other uses of ratios and percentages are considered. Finally, because rates are a special type of ratio, rates and also percentages of change will be treated briefly.

DESCRIPTION OF THE DISTRIBUTION OF NONQUANTITATIVE CHARACTERISTICS

The Problem. The simplest case of a nonquantitative characteristic is that of a dichotomous characteristic with regard to which every individual unit in a series is enumerated as being in one of two mutually exclusive categories. As an example, let us consider the characteristic sex, with regard to which a number of individuals are enumerated as being either male or female. In statistical analysis it is sometimes convenient to think of the two categories (male and female) of a dichotomous nonquantitative characteristic (sex) as a division of individuals into those who possess an attribute (maleness) and those who do not possess the attribute. The process of gathering data on the sex distribution of a number of individuals consists of enumerating each individual as either possessing the attribute "maleness" or not possessing the attribute "maleness," or stated slightly differently, of enumerating each individual as being either "male" or "not-male." When a series of individuals have been so enumerated, as is seen from the classification of characteristics on page 66, the data on the distribution of the attribute "maleness" for the series consists of two numbers, those who possess it and those who

71

do not. Given these data, the problem of description of the distribution becomes one of combining the two numbers in some way so as to get one number, a summarizing measure, which will describe for the series or group of individuals their sex distribution. The types of summarizing measures most frequently employed for this sort of description are ratios, proportions, and percentages.

Ratios. A ratio is an indicated or actual quotient which denotes the relation in size of one number to another. In the particular type of ratio in which we are interested here the two numbers are the frequencies of two enumerations. In the example of sex distribution let us suppose that the data secured by enumeration of the students in a statistics class with regard to sex are 12 males and 8 not-males (females). The first summarizing measure we shall compute is the ratio of males to females, which may be written in any of the following equivalent forms:

$$12{:}8 = \frac{12}{8} = 3{:}2 = 1.5$$

The first three ways of writing the ratio are "indicated" quotients; the last way is an "actual" quotient since a division has actually been made. In our example we can use any of these forms of the ratio to describe the sex composition of the class; thus, "The ratio of males to females is three to two," or, "The ratio of males to females is 1.5." The actual quotient is the most condensed form, of course.

Let us put the definition of the type of ratio used as a summarizing measure of the distribution of an attribute into symbols so that in any similar case we can substitute the data obtained from enumeration into a formula and get the required summarizing measure. For a dichotomous nonquantitative characteristic let us denote the attribute specified by the major category as A, the number of individuals who possess that attribute as (A), and the number of individuals who do not possess that attribute as (α). Then the ratio describing the relative frequency of the major category to the minor category is

$$\text{Ratio} = \frac{(A)}{(\alpha)} \tag{1}$$

Note that such ratios are not dependent on the absolute number of individuals enumerated, and, therefore, they are relative measures which can be compared with similarly computed measures for other groups of individuals. Suppose that in the graduate school of the university enumeration shows that for all graduate students there are 351 males and 234 females. If we wish to compare the sex composition of the statistics class with that of the entire graduate school we can substitute the graduate school data in formula (1) and obtain as ratio of males to females,

$$\frac{351}{234} = 1.5$$

The ratio is identical with the corresponding ratio for the statistics class, and we can, therefore, say that the sex composition, or the sex distribution, of the statistics class is identical with that of the graduate school, although the absolute frequencies in the two groups are quite different. The utility of summarizing measures such as ratios is evident when comparisons are to be made.

Modified ratios. If a ratio is greater than one, the most common way of writing it is to give the actual quotient—1.5 in the above case. Whenever a single number like this is given as a ratio, the number "1" is understood as its denominator. When a ratio, or certain ratios in a series of ratios, is likely to be less than one, a device of multiplying the ratio by 100 or 1,000 is often employed to avoid expressing the ratio as a decimal fraction less than one. The result is not a ratio in the strictest sense, but is actually a ratio multiplied by some power of ten. Since such measures are usually called "ratios," we shall do likewise, but in this chapter shall enclose the term in quotation marks when used in this sense to distinguish it from the stricter sense of ratio. For instance, for cities, states, or other population units the usual practice is to compute the number of males per 100 females as the "sex ratio" rather than the number of males per female. The "sex ratio" of the statistics class put into the conventional form becomes 150 rather than 1.5. When an arbitrary base like 100 females is used in a modified ratio, the base must always be stated.

Conventions regarding percentages. Since percentages are so frequently employed, certain conventions have become established regarding their use. Some of these are listed below as suggestions in the computation and presentation of percentages.

1. Unless there is reason for doing otherwise, percentages are usually given to one decimal place. Occasionally more or less accuracy may be desired or justified. Carrying a percentage to one decimal place means that in actual calculation the division is carried to four decimal places. The decimal point is then moved two places to the right to multiply by the base 100, leaving two digits to the right of the decimal. Finally, the rightmost digit is dropped in the process of rounding to one decimal place.

2. The process of rounding is done as follows: if the rightmost digit is 0, 1, 2, 3, or 4, it is dropped with no change made in the digit next to the left; if the rightmost digit is 6, 7, 8, or 9, it is dropped and the digit next to the left is raised one unit; if the rightmost digit is 5, and there is a remainder (in the process of division), the digit next to the left is raised one unit; if the rightmost digit is 5, and there is no remainder, the digit next to the left is raised one unit if it is odd, or is unchanged if it is even.

3. In the construction of a table which includes percentages, the bases of

the percentages should be so clearly indicated that there is no ambiguity possible. A device for indicating the base is to have one row or column assigned to the total or base, with only 100.0 percent entries in it.

4. When all of the components of a total are expressed as percentages of the total, they should be checked by adding, since their sum should be 100.0 percent. Frequently the sum will be 99.9 or 100.1 due to inaccuracies involved in carrying the percentages to only one decimal place. In such a case, the entry opposite the total is always written as 100.0 percent, but there are three choices of procedure about the component percentages: (a) the component percentages may be left as they are; (b) the largest component percentage may be increased or decreased by 0.1 to make the sum check; (c) the percentage or percentages which were changed most in the process of rounding may be changed by 0.1 in a direction different from that of the change made in rounding. If the sum of the percentages varies more than 0.1 or 0.2 from 100.0, one should check carefully all processes to see if the discrepancy is caused by a mistake.

5. In any given study one usually sets a lower limit, such as 100 or 50, below which percentages are not computed.[1]

Relationship between order of units and type of distribution. In any problem involving statistical analysis it is necessary to distinguish between different orders of varying units. For example, the varying units on which observations as to marital status were made in 1949, and about which data are recorded in Table 6, are individual human beings. When summarizing measures are computed for a *group* of individuals, the summarizing measures refer to the group. Thus, we see that a characteristic may be nonquantitative when related to an individual but quantitative when it refers to a group. For the individual marital status is nonquantitative since the individual falls into one of several categories. In any group, however, the percentage who are single (or who are in any of the other categories) can theoretically take any value from zero to 100 percent. Thus, marital status (or the percentage in any particular status) becomes a quantitative characteristic when groups are the varying units.

Proportions. It will be remembered that a ratio is defined as the quotient of two numbers, and the type of ratio we have just illustrated is the quotient of two component frequencies. Another type of ratio of great importance is a *proportion* where the numerator of the quotient is

---

[1] Not all statisticians observe this convention. The reason for the convention seems to be the generalizing function implied in the use of percentages. Literally the word means so many per one hundred, with a strong suggestion of describing at least a hundred units or more. However, as the differentiation between the descriptive and the generalizing functions of statistics becomes better recognized, it is possible that as purely descriptive measures of the distribution of a characteristic among a unique group of units, percentages computed on small bases will become more generally permissible. If, of course, percentage composition of a universe is being estimated from a sample, confidence limits or some other measures of precision presented in Part III should be used to indicate the reliability of the generalization.

the frequency of one category and the denominator of the quotient is the total number of units enumerated. As before, we let the symbols,

$$(A) = \text{number of units possessing attribute } A$$
$$\text{and } (\alpha) = \text{number of units not possessing attribute } A$$

Then the total number of units enumerated, which we shall designate as N, is

$$N = (A) + (\alpha) \tag{2}$$

In the case of the statistics class substitution in (2) gives as the total number of students,

$$N = 12 + 8 = 20$$

In the group enumerated the proportion of units possessing the attribute A, which we shall designate as p, is found by the relation,

$$p = \frac{(A)}{N} \tag{3}$$

In the statistics class the proportion of males in the class is found by substituting in (3), thus,

$$p = \frac{12}{20} = .6$$

The proportion complementary to p—that is, the proportion of units not possessing attribute A—is designated as q and is obtained by the relation,

$$q = \frac{(\alpha)}{N} \tag{4}$$

In the statistics class q is the proportion of females in the class and is obtained by substitution in (4), thus,

$$q = \frac{8}{20} = .4$$

When we say that two proportions are complementary, we mean that their sum is one. Thus for any attribute A, which all of N units either possess or do not possess, when p and q are defined as above, the relation,

$$p + q = 1 \tag{5}$$

is always true, as is illustrated in the statistics class by the fact that

$$.6 + .4 = 1$$

Use of proportions. Proportions are used as summarizing measures of the distribution of nonquantitative characteristics as alternatives to ratios of component frequencies. Instead of using the "sex ratio" to de-

scribe the sex distribution of the population of the Southeast in 1950, we can use the proportion of males, computed by substituting in formula (3), thus,

$$p = \frac{74,633,000}{150,697,000} = .4953$$

Along with this proportion, if we choose, we can also use the proportion of females, which we obtain either by substituting in formula (4), thus,

$$q = \frac{76,064,000}{150,697,000} = .5047$$

or by substituting the value of p in formula (5), thus,

$$.4953 + q = 1$$
$$q = 1 - .4953 = .5047$$

When the nonquantitative characteristic, whose distribution is being described, has only two categories as in the sex illustration, it does not usually matter whether one uses a ratio of component frequencies or one or two proportions to describe the distribution. When the characteristic has more than two categories, however, the use of a series of proportions is usually more efficient than a series of ratios of component frequencies in conveying the description.

As an example, let us consider the distribution of the nonquantitative characteristic, "marital status," among all the females 14 years of age and over in the United States in 1949. Table 6 shows data obtained from the Current Population Reports in which persons 14 years of age and over were classified as being in one of five categories of marital status.

Table 6. DISTRIBUTION OF PERSONS FOURTEEN YEARS OF AGE AND OVER BY SEX AND MARITAL STATUS, UNITED STATES, APRIL, 1949.

[Estimates for 1949 are rounded to the nearest thousand without adjustment to group totals, which are independently rounded.]

Marital status	Males	Females
All marital conditions............	53,448,000	56,001,000
Single.......................	13,952,000	11,174,000
Married, spouse present........	35,323,000	35,323,000
Married, spouse absent.........	1,151,000	1,690,000
Widowed....................	2,181,000	6,582,000
Divorced...................	842,000	1,233,000

Source: Bureau of the Census, Current Population Reports. Population Characteristics. Series P-20, No. 25, "Changes in Number of Households and in Marital Status: 1940–1949." Table 2.

Now if we were to try to describe the distribution of the nonquantitative characteristic, "marital status," with its five categories, by ratios between each pair of component frequencies, we should have to compute 10 ratios, for there are 10 possible combinations of component frequencies by pairs. For instance, we might form the ratio of married females, spouse present, to single females, thus,

$$\frac{35,323,000}{11,174,000} = 3.16$$

A better way to describe the distribution is to compute the proportion the frequency of each category is of the total number. For females these proportions are computed as follows by successive substitutions in formula (3),

$$\text{Proportion of females who are single} = \frac{11,174,000}{56,001,000} = .1995$$

$$\text{Proportion of females who are married, spouse present} = \frac{35,323,000}{56,001,000} = .6308$$

$$\text{Proportion of females who are married, spouse absent} = \frac{1,690,000}{56,001,000} = .0302$$

and similarly for proportions who are widowed and divorced, getting for these groups proportions of .1175 and .0220, respectively. This series of five proportions describes the distribution of marital status among the females 14 years of age and over in the United States in 1949. They may be compared with one another to show the relative frequencies of the various marital conditions. The greatest advantage of these proportions, however, lies in their utility for comparing this group with another with regard to distribution of marital status.

Percentages.   Just as ratios of component frequencies are often multiplied by a power of 10, so also are proportions. When a proportion is multiplied by 100, it is called a percentage. While proportions are preferred as summarizing measures by statisticians if they are to be used in further analysis, percentages are usually preferred by the reading public. Therefore, while the statistician may use proportions in the stage of analysis, he usually changes them into percentages before presenting them in the tables, charts, or text of a report. The change is made by multiplying the proportion by 100, which involves simply moving the decimal point two places to the right. Percentages can be determined directly by the following relation,

$$\text{Component percentage} = \frac{f}{N} \times 100 \qquad (6)$$

where f = frequency of a class
and N = total frequency

By moving the decimal point two places to the right in the proportions computed for females, and by substituting successively in formula (6) the data on males from Table 6, the two series of percentages shown in columns (1) and (2) of Table 7 are obtained.

Table 7. PERCENTAGE DISTRIBUTION OF MALES AND FEMALES 14 YEARS OF AGE AND OVER BY MARITAL STATUS, UNITED STATES, 1949.

Marital status	Males	Females
All marital conditions.............	100.0	100.0
Single........................	26.1	19.9
Married, spouse present.........	66.0	63.1
Married, spouse absent..........	2.2	3.0
Widowed......................	4.1	11.8
Divorced.....................	1.6	2.2

Source: Table 6.

Therefore, for further analysis of series of these summarizing measures (ratios, proportions, and percentages) we shall need to use the body of methods which describes the distributions of quantitative characteristics. These methods will be presented in the next chapter.

OTHER USES OF RATIOS AND PERCENTAGES

Ratios. In addition to their use for describing the distributions of nonquantitative characteristics among a series of units, ratios are used for the purpose of comparing any two frequencies or any two quantities. A modified "ratio" with a base of 1,000 which is frequently met in population literature is the "fertility ratio," defined as the number of children under 5 years of age per 1,000 women aged 20–44 (or alternate age ranges that approximately encompass the childbearing period, such as 15–49). As a formula,

$$\text{Fertility ratio} = \frac{\text{Number of children under 5}}{\text{Number of women aged 20–44}} \times 1,000 \qquad (7)$$

The "fertility ratio" of the United States for 1950 can be computed from the data of Table 4. The number of children under 5, both male and female, is 16,324,000, while the number of females 20–44 is 28,651,000.

$$\text{Fertility ratio} = \frac{16,324,000}{28,651,000} \times 1,000 = 569.8$$

Since 569.8 is not a ratio in the strictest sense, its definition, "number of children under 5 per 1,000 women aged 20–44" should always be explicitly stated.

Ratios may involve different sorts of units, such as "population per square mile"—a ratio of people to square miles. Only in the case of a ratio composed of two additive components of a total can we consider the ratio as a "pure" number which can be written without designating units.

RATES

Description of change. Sociologists are interested not only in describing the distributions of various characteristics at any one point of time, but also in describing changes in distributions of characteristics. There are several aspects of change which may be described. Some aspects require the use of time series analysis (to be treated in Chapter 11) and other more elaborate statistical methods than will be treated in this chapter. Certain simple aspects of change, however, can be described by summarizing measures which are special types of those we have already defined in this chapter.

Rates. A rate is a special type of "ratio" where the numerator is the number of some particular type of events which occur during a unit period of time, the denominator is the number of some type of units (which may also be events) to which the occurring events are related, and the quotient is usually multiplied by 100 or 1,000. Rates always involve the concept of change, and therefore also of time, since change has to be measured over a period of time. Thus, rates are dynamic as distinguished from static component ratios or percentages which describe composition with regard to some characteristic at a single point of time.

Birth and death rates. Birth and death rates are probably the most important rates for students of population. They relate the number of births or deaths in a given year to the midyear population. The crude death rate is the total number of deaths occurring in an area during a year per 1,000 midyear population and can be computed by the following formula:

$$\text{Crude death rate} = \frac{\begin{array}{c}\text{Number of deaths occurring in the}\\\text{area during a given year}\end{array}}{\begin{array}{c}\text{Number of people living in the area}\\\text{at the midpoint of that year}\end{array}} \times 1,000 \quad (8)$$

The formula for the crude birth rate is similar.

$$\text{Crude birth rate} = \frac{\substack{\text{Number of births occurring in the} \\ \text{area during a given year}}}{\substack{\text{Number of people living in the area} \\ \text{at the midpoint of that year}}} \times 1,000 \qquad (9)$$

These rates are called *crude* because they deal with events related to the entire population. If only a specific portion of the population is considered, the rates are called specific rates. The formula for an age-sex specific death rate (for example, males 20–24) follows:

$$\substack{\text{Age-sex specific} \\ \text{death rate}} = \frac{\substack{\text{Number of deaths in age-sex} \\ \text{group during given year}}}{\substack{\text{Number of people in age-sex} \\ \text{group at midpoint of year}}} = 1,000 \qquad (10)$$

Another important rate for sociologists and demographers is the rate of natural increase.

$$\text{Rate of natural increase} = \frac{\substack{\text{Excess of births over} \\ \text{deaths during year}}}{\text{Midyear population}} \times 1,000 \qquad (11)$$

In computing rates it is conventional for the denominator to be the average or midyear frequency. When we compute a rate of change in which the denominator is the frequency or amount present at the beginning of the period, it is conventional to multiply it by 100 and call it the "percentage change." For example, the population of North Carolina in 1940 was 3,571,623, and in 1950 it was 4,061,929. The formula for percentage change in a frequency for a unit period of time is as follows:

$$\text{Percentage change} = \frac{f_2 - f_1}{f_1} \times 100 = \left(\frac{f_2}{f_1} - 1 \right) 100 \qquad (12)$$

where f_1 = initial frequency at beginning of unit period
and f_2 = frequency at end of unit period

Substitution of the population figures for North Carolina gives

$$\text{Percentage change} = \frac{4,061,929 - 3,571,623}{3,571,623} \times 100 = 13.7 \text{ percent}$$

This is obviously a percentage increase, but if the result were a negative number rather than a positive number, it is a percentage decrease. Percentages of change may relate to phenomena other than frequencies but the formula is the same as (12) with quantities substituted for frequencies.

Sometimes the term rate is applied to a "hybrid" sort of measure. The "labor force participation rate" is an example of this sort of measure.

This "rate" indicates the proportion of the population of a specified age-sex group who are in the labor force. Since labor force participation is currently defined in the United States on the basis of activity or status during the week preceding the survey, the "rate" reflects the proportion of the population who have manifested a certain type of activity recently in contrast with the proportions relating to more static characteristics, such as sex or race. The rate is defined as follows:

$$\text{Labor force participation rate} = \frac{\text{Number of persons in the age-sex group classified as being in the labor force in the preceding week}}{\text{Number of persons in age-sex group at specified time}} \quad (13)$$

For males aged 20–24 years in April 1950 the labor force participation rate was

$$\frac{4,490,000}{5,457,000} = .823$$

SUGGESTED READINGS

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chap. VII.

Peatman, John Gray, *Descriptive and Sampling Statistics* (New York: Harper, 1947), Chap. 3.

Zeisel, Hans, *Say It With Figures* (New York: Harper, 1947), Chaps. IV and V.
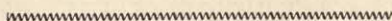
Quantitative Distributions: Measures of Central Tendency

SINCE the observed measures of the incidence of a quantitative characteristic for a series of units can take many values—often a theoretically infinite number of different values—the problem of analysis and description of a series of such varying measures is more complicated than in the case of a series of measures of the incidence of a nonquantitative characteristic, where the measures can take only two values—"all" or "none." Therefore, we pause for a preview of the methods of this and the following chapter and for some comments on the range of their utility.

Organization. In this chapter, following the introductory material, will be found a treatment of the methods of describing the distribution of a quantitative characteristic among a series of varying units by techniques for presenting the distribution as a whole. Next, three aspects of quantitative distributions will be considered with the methods of investigating, describing, and presenting each. These three aspects are (1) the central tendency of the distribution, (2) the dispersion of the distribution, and (3) the form of the distribution. The first of these is discussed in this chapter, the other two in Chapter 9. For describing the distribution as a whole, we shall use arrays of ordered but ungrouped data, frequency tables of data grouped into class intervals, modifications of frequency tables, such as cumulative frequency tables and percentage distribution tables, and the corresponding graphic forms of representation of quantitative distributions. For investigating and describing the central tendency of the distribution, we shall use summarizing measures—the arithmetic mean, the median, and the mode—which are all types of averages. Since any of these measures of central tendency can be expressed as one number, there is no need for separate graphic representation of them, although they can be represented on the charts already made of the distribution. For describing the dispersion of the distribution, we shall also use summarizing measures—the range, the mean deviation,

the quartile deviation, the 10–90 percentile range, and the standard deviation. For describing the form of the frequency distribution, we shall use a classification of plotted distributions by type of form and a measure of skewness.

Utility of the methods of this chapter. As to the range of applicability of these methods, they are limited only by the nonmeasurability of certain phenomena in which sociologists are interested, for data about any measurable characteristic distributed unequally among a series of units can be treated by these methods. Let us point out that we are using the word "characteristic" in its broadest sense, not in the more limited sense which denotes a personal character trait. A sampling of what we mean by measurable characteristics is the following list of characteristics grouped under headings which indicate the types of individual units for which the listed characteristic might be measured in a sociological research project.

Individual person	*Public welfare agency*	*Farm*
Age	Number of staff members	Years of tenure of operator
Last grade completed	Average number of years	ator
Number of children	of professional training	Gross income during a
Weekly income	of staff members	specified year
Number of organizations participated in during specified time	Number of active cases	Number of acres cultivated
	Average of amount of old-age assistance grants	Number of cows

Such a list might be continued almost indefinitely by extending both the list of types of units one might study and the list of characteristics one might measure for each type of unit. If for each of a group of the units listed as headings, observations are made on some one of the characteristics listed and are recorded in numerical form, these numerals representing the measures of the characteristic constitute the data on the distribution of that characteristic among the group studied and are ready to be analyzed by the methods of this chapter. Since the values of such measures vary, either the varying measures themselves or the characteristic they measure may be called a "variable."

All measures of quantitative characteristics will not be obtained in the same way and will not be of the same nature. For instance, the measures of the characteristic "number of children" for a person can be only in whole units (children) since the characteristic increases in unit jumps; the measures of income are usually expressed in hundredths of units (dollars) since cents are the smallest increment by which this characteristic can increase; while the characteristic age, although usually observed and recorded only to the last or nearest whole unit (year), can be measured to almost any fraction of a unit since the characteristic in-

creases continuously. Quantitative variables formed by the measures of characteristics which increase by jumps are called *discrete* variables; while those formed from the measures of characteristics which increase gradually, with the possibility of taking every value in an interval of a scale, are called *continuous* variables.

DESCRIPTION OF THE DISTRIBUTION AS A WHOLE

Unordered data. It is difficult to comprehend or to picture the distribution of a measurable characteristic from the records of a great number of observations until they have been put in some sort of order. It is even difficult to tell much about distributions from unordered data on moderate sized groups. We illustrate this by showing data relating to "number of children borne" from a study of 117 white tenant farm women. The number of children borne was first recorded along with other information for each woman on an individual collection form. When the data were assembled onto one assembly sheet, the actual number of children borne by each woman was written in the column labeled for brevity, "number of children," and in the row assigned to that woman. Table 8 shows the way these numbers of children borne appeared on the assembly sheet where they were listed in the order of the serial numbers of the individual collection forms.

From such a compilation of records it is possible to get some idea of the distribution of the characteristic "number of children borne." (The "number of children borne" refers to the number of live children borne, whether or not they are living now, and in this problem we shall call this characteristic "fertility," although other studies may use differently defined measures of the characteristic "fertility.") From scanning the columns listing the number of children, one receives the impression that for 1938 these are a group of rather high fertility, yet as long as the data are left in this form there is no way to place this group as to level of fertility (or average number of children borne per woman) or to tell with much precision how greatly the individual women vary from such an average. Because this is such a small group, one can pick out the smallest and greatest numbers of children borne, 0 and 16, and thus determine the range of variation; but if there were a much greater number of records, even this would be difficult.

The array. A number of simple methods which would order and condense these data may occur to the reader. Perhaps the least mathematical of these is to arrange the two figures for each woman (serial number and number of children borne) in order of the number of children borne instead of in order of the serial number as they are shown in Table 8. They are so arranged in Table 9, in order of the measures, number of

Table 8. EXCERPTS FROM ASSEMBLY SHEET SHOWING SERIAL NUMBER AND NUMBER OF CHILDREN BORNE FOR 117 WHITE TENANT FARM WOMEN, 1938

Serial number *	Number of children	Serial number	Number of children	Serial number	Number of children	Serial number	Number of children
1	7	35	4	68	6	100	5
2	9	36	13	69	6	101	6
3	4	37	11	70	11	102	0
4	16	39	8	71	4	103	12
5	10	40	7	72	10	104	6
6	7	41	8	73	12	105	2
7	10	42	2	74	8	106	3
8	5	43	1	75	3	107	6
9	13	44	11	76	1	108	4
12	6	45	4	77	11	109	7
13	9	46	2	80	6	110	6
14	4	47	6	81	8	111	1
15	2	48	4	82	6	112	1
16	6	49	8	83	11	113	5
17	10	50	2	84	11	114	4
18	13	51	4	85	12	115	7
19	5	52	7	86	2	118	9
20	6	53	13	87	6	120	4
21	2	54	4	88	2	121	8
22	11	55	1	89	8	122	9
23	3	56	7	90	5	123	9
24	9	57	1	91	12	124	8
25	6	58	4	92	3	125	7
26	5	59	5	93	7	126	3
27	5	60	7	94	5	127	4
28	4	63	8	95	12	128	5
29	7	64	8	96	10	129	2
30	1	65	6	97	6		
31	2	66	7	98	7		
32	1	67	10	99	7		

* Only 117 of 129 records are listed here; 10 of the other 12 records are of owners' wives and 2 are incomplete.

Source: Field study by Margaret Jarman Hagood.

children borne, from highest to lowest. Such an arrangement is called an array, whether the order is from highest to lowest or from lowest to highest. The values that are arrayed are the measures of the individual women with regard to the characteristic studied. The serial numbers of

Table 9. ARRAY OF THE NUMBERS OF CHILDREN BORNE FOR 117 WHITE TENANT FARM WOMEN, 1938

Serial number	Number of children	Rank	Serial number	Number of children	Rank
4	16	1	39	8	34.5
			41	8	34.5
9	13	3.5	49	8	34.5
18	13	3.5	63	8	34.5
36	13	3.5	64	8	34.5
53	13	3.5	74	8	34.5
			81	8	34.5
73	12	8	89	8	34.5
85	12	8	121	8	34.5
91	12	8	124	8	34.5
95	12	8			
103	12	8	1	7	46.5
			6	7	46.5
22	11	14	29	7	46.5
37	11	14	40	7	46.5
44	11	14	52	7	46.5
70	11	14	56	7	46.5
77	11	14	60	7	46.5
83	11	14	66	7	46.5
84	11	14	93	7	46.5
			98	7	46.5
5	10	20.5	99	7	46.5
7	10	20.5	109	7	46.5
17	10	20.5	115	7	46.5
67	10	20.5	125	7	46.5
72	10	20.5			
96	10	20.5	12	6	61.5
			16	6	61.5
2	9	26.5	20	6	61.5
13	9	26.5	25	6	61.5
24	9	26.5	47	6	61.5
118	9	26.5	65	6	61.5
122	9	26.5			
123	9	26.5			

Source: Table 8.

the women are included merely for identification and are not an essential part of the array. The column of ranks will be treated in the next section.

With the measures ordered into an array, the features of the distribution of fertility among this group become much clearer. One can see that six children is about the average number for these women, that the measures are rather well scattered over the range from 0 to 16, with fewer cases at the ends of the range than in the middle. For two reasons it is possible

Table 9. ARRAY OF THE NUMBERS OF CHILDREN BORNE FOR 117 WHITE TENANT FARM WOMEN, 1938—(Continued)

Serial number	Number of children	Rank	Serial number	Number of children	Rank
68	6	61.5	108	4	86.5
69	6	61.5	114	4	86.5
80	6	61.5	120	4	86.5
82	6	61.5	127	4	86.5
87	6	61.5			
97	6	61.5	23	3	96
101	6	61.5	75	3	96
104	6	61.5	92	3	96
107	6	61.5	106	3	96
110	6	61.5	126	3	96
8	5	74.5	15	2	103.5
19	5	74.5	21	2	103.5
26	5	74.5	31	2	103.5
27	5	74.5	42	2	103.5
59	5	74.5	46	2	103.5
90	5	74.5	50	2	103.5
94	5	74.5	86	2	103.5
100	5	74.5	88	2	103.5
113	5	74.5	105	2	103.5
128	5	74.5	129	2	103.5
3	4	86.5	30	1	112.5
14	4	86.5	32	1	112.5
28	4	86.5	43	1	112.5
35	4	86.5	55	1	112.5
45	4	86.5	57	1	112.5
48	4	86.5	76	1	112.5
51	4	86.5	111	1	112.5
54	4	86.5	112	1	112.5
58	4	86.5			
71	4	86.5	102	0	117

to comprehend more about the features of the distribution of the characteristic from this array than from many other arrays: first, the number of measures is only 117, which means there is no great mass of data to confuse one; second, the measures take only 15 different values, so that this array is almost the equivalent of a frequency table. Yet one can well imagine that if there were 500 or 1,000 measures arrayed, with almost every one of them having a different value, that an array would be less informative than in the present case and that we should need to condense the material further. In fact, if the number of cases is much greater than

in this example, it becomes cumbersome and tedious to arrange an array, at the same time the array conveys less information, especially if the measures take many different values. Therefore, the making of an array is often dispensed with since the most important parts of the further analysis are not dependent upon the array.

Ranking. Often arrays may include an additional column listing the rank of each measure, as does Table 9. In any array the measure with the highest numerical value has the rank "1," the measure with the next highest value has the rank "2," and so on to the measure with the lowest value, which has a rank equal to the number of measures arrayed, 117 in this case. Although arrays may be arranged in either ascending or descending order, descending is preferable, particularly for those just learning the conventions, because in descending arrays the first measure is ranked "1" and there is less likely to be confusion in the designation of ranks, or of other measures based on ranks, such as quartiles.

It is clear that the woman with serial number 4, who has borne 16 children, ranks first [1] in fertility among this group, and that the woman with serial number 102, who has borne no children, ranks lowest, or one hundred and seventeenth. But the assignment of ranks to the other women, or rather to their measures in fertility, is not so simple. Shall we assign rank "2" to the woman with serial number 9, 18, 36, or 53? Each of them has 13 children. There are several ways of handling the assignment of ranks in cases of ties, but the following procedure is advised. Evidently this group of four women should get ranks "2," "3," "4," and "5," and yet they should all get the same rank because they all have the same number of children. Therefore, we add the ranks, "2," "3," "4," and "5," and divide the sum by 4 to get an average rank, which is then assigned to each member of the group. Thus,

$$\frac{2 + 3 + 4 + 5}{4} = \frac{14}{4} = 3.5$$

is the rank assigned to each woman having 13 children. Now as we go on to the next group in assigning ranks, we must remember that we should start with the next integer, that is, with "6," because we have used up to and through rank "5." For the group of women who have 12 children, then, the average rank which will be assigned each member of the group is

$$\frac{6 + 7 + 8 + 9 + 10}{5} = \frac{40}{5} = 8$$

---

[1] Just as we often refer descriptions of distributions of characteristics to the group of units manifesting the characteristic, so do we refer particular parts of the description of measures to the units manifesting those degrees of the characteristic. Although it is actually the measures that are ranked in order of their numerical value, we speak of the woman whose measure has a certain rank as herself having that rank.

The process of computing ranks in cases of multiple ties can actually be done without addition. If the number tying for a place is odd, as in the second case illustrated, simply take the middle one of the ranks involved. Thus from

$$6, 7, 8, 9, 10$$

choose "8" as the average rank to be assigned to each of the group. If the number tying for a place is even, as in the first case illustrated, take the midpoint between the two middle ranks involved. Thus with

$$2, 3, 4, 5$$

choose the midpoint between "3" and "4," that is, "3.5," as the average rank to be assigned to each of the group.

The particular circumstance which gives rise to so many ties in rank in this example is the fact that the number of values these measures of fertility can take is quite limited; it is impossible to have fractional values on the measure, "number of children borne," and it is very rarely that a woman has more than 16 children. Since no one of these women happens to have 14 or 15 children, only 15 possible values remain (including zero) for these 117 measures to take, and hence the values are necessarily bunched, causing ties.

Grouping data into class intervals. Whether or not an array is made of the data on a quantitative distribution, the most commonly used arrangement for further analysis or for presentation is a frequency table. Before a frequency table can be constructed, however, the data must be grouped into class intervals. And before grouping can actually be done, the class intervals must be chosen. There are no definite rules for setting up class intervals, but they should be chosen in such a way as to fulfill the following conditions as nearly as is possible.

1. The number of class intervals should be great enough to avoid sacrificing too much of the accuracy gained by the precision of the observations.

2. The number of class intervals should be small enough to avoid many vacant classes or too great fluctuations in frequencies of adjoining classes.

3. The class intervals should be equal if possible, since this makes further analysis easier; if, however, there are a few widely divergent measures, a compromise is to have all class intervals equal except one or both of the end classes and to have these "open," indicated by some such expression as "16 and over." These open class intervals cause problems, though, and should be avoided if at all practicable.

4. The class intervals should be spaced so as to have any irregular points of concentration of values fall at their midvalues. This is necessary because the analysis of grouped data proceeds upon the assumption that the mean value of all the measures in the class interval falls at the midvalue of the interval.

If there are heavy weightings at either end of the class interval, this assumption will not be justified.

5. The class intervals should cover every possible value of the measures within the range of values observed, but should not be overlapping.

6. The setting up of the class limits and the specifying of the class intervals should be done with consideration of three features of the data: first, the discreteness or continuity of the numerical values of the characteristic being studied; second, the precision of measurement of the characteristic in the gathering of data; third, the conventions observed in recording the data. (Age, for instance, is usually recorded differently from most other measures.) One must differentiate carefully between the *true* limits of a class interval, which may or may not be included in the interval, and the *specified* limits, which (to avoid simultaneously overlapping and awkward phrasing) may not be identical with the true limits.

In our example the class intervals immediately suggested by the material are those of "no children," "one child," "two children," and so on. Let us see how these class intervals meet the criteria just proposed. (1) No accuracy of observation is sacrificed at all by this choice of intervals. (2) We shall test this criterion with a graphic presentation of the frequency distribution constructed with these intervals later, and shall find that there are too many classes. (3) The classes are all equal if we consider every possible number of children a woman may have borne as an "equal" class. That is, the classes are equal so long as we treat the discrete variable, "number of children borne," as a discrete variable which can take only integral values. When we assume continuity of the variable, as will be explained under (6), we shall find that the zero interval offers a special problem. (4) There are no irregular points of concentration to be considered in this distribution. (5) The classes cover all values of the measure which can possibly be observed within the range of observation, and the intervals do not overlap.

(6) The considerations involved in this criterion require a more lengthy exposition. Since the observed measures of the characteristic can take only integral values (whole numbers) and since the measures have been recorded with complete precision as integers, there is no problem here in the grouping of observed measures into intervals; but there is a problem in setting up the class limits and in specifying the class intervals, which will be used in the further analysis of the grouped data. Let us anticipate for a moment the process of finding the arithmetic mean, which is 6.3 children, and see how we are going to interpret it. In everyday situations 6.3 children has no meaning. Yet, such abstractions as 6.3 children are frequently found useful in condensing information about groups. It might be quite meaningful to differentiate between a group with a mean number of children of 6.4 and one with a mean of 6.1. Summarizing measures, which may take fractional values, are *abstractions*,

but we shall need to refer them to the same intervals as those into which the *real* observed measures are grouped. Keeping in mind this need for referring fractional values to class intervals, we must set up our class limits *as if* the variable formed by measures of the characteristic were continuous. That is, the limits of the classes must be set so that they will include not only every value which may be observed, but also every value which may be employed as an abstraction, since in the processes of computation these values will have to be referred to class intervals. The obvious limits to set up are the halfway points between integers—4.5–5.5, 5.5–6.5, etc. We shall call these the "true" limits; but they are not satisfactory to use in specifying the class intervals because they indicate that the intervals overlap, the midpoints between integers being included in two intervals, and thus violating criterion (5). It is customary in the definition of a class interval to define each class interval as including its lower limit but as not including its upper limit (although occasionally this convention is reversed). The precise way of specifying the interval 4.5–5.5 is "4.5 up to but not including 5.5." Let us emphasize, however, that in such a definition 5.5 is still the upper limit of the interval, since a value of the measure differing only an infinitesimally small amount from 5.5 will be included in the interval. The specification of the interval by the phrase "4.5 up to but not including 5.5" is too awkward to use in tables, and, therefore, we specify its upper limit by the next smaller value that can be shown with the number of decimal places being used. Thus, the interval is specified as "4.5–5.4" though we must keep in mind the fact that 5.5 is the "true" upper limit.

When we have occasion to use class limits in analysis, for example, in determining the midpoint of the class interval, we must always use the "true" limits. The relationship between the "true" limits and the "score" limits (as the specified limits are called) depends upon the nature of the data. For discrete data the lower "true" limit is generally identical with the lower "score" limit. However, the upper "true" limit of a class is generally identical with the lower "score" limit of the next class. This same set of conventions is followed in setting up class limits for age data where the age is "age at last birthday." When the data are measured or derived, however, the true lower limit of a class lies halfway between the lower score limit of the class and the upper score limit of the previous class. Similarly, the upper true limit lies halfway between the upper score limit of the class and the lower score limit of the next class. The logic of this relationship follows from the fact that the measured or derived figures were rounded up or down depending upon which side of the true limit they fell.

Illustrations of true limits, score limits, and midpoints can be seen in Table 10. Note that the upper true limit of one class is always the same

as the lower true limit of the next class and that the midpoint of a class is the midpoint of the true limits.

Returning to *number of children borne* we must now treat the troublesome interval containing zero. In getting the midpoint of the class it offers no problem, for we merely write zero as a possible number of children a woman might have borne. However, when we wish to flatten out zero a half unit in either direction to specify the true class limits, we obtain the class limits −0.5–0.5. There is a difference between the abstrac-

Table 10. THREE EXAMPLES OF TRUE LIMITS, SCORE LIMITS, AND MIDPOINTS

Number of children borne			Age at last birthday			Percent of population urban		
True limits	Score limits	Mid-points	True limits	Score limits	Mid-points	True limits	Score limits	Mid-points
0.5–1.5	0.5–1.4	1	0–5	0–4	2.5	0.00–9.95	0.0–9.9	4.975
1.5–2.5	1.5–2.4	2	5–10	5–9	7.5	9.95–19.95	10.0–19.9	14.95
2.5–3.5	2.5–3.4	3	10–15	10–14	12.5	19.95–29.95	20.0–29.9	24.95
3.5–4.5	3.5–4.4	4	15–20	15–19	17.5	29.95–39.95	30.0–39.9	34.95
4.5–5.5	4.5–5.4	5	20–25	20–24	22.5	39.95–49.95	40.0–49.9	44.95
5.5–6.5	5.5–6.4	6	25–30	25–29	27.5	49.95–59.95	50.0–59.9	54.95
6.5–7.5	6.5–7.4	7	30–35	30–34	32.5	59.95–69.95	60.0–69.9	64.95
etc.			etc.			etc.		

tion of a fraction of a child and the absurdity of a negative half child. Even statisticians balk at the latter concept and will not represent it in tables or in charts, although they do imply an acceptance of it later in treating the class interval containing zero as of a width equal to those of the other classes. They will not, however, represent this vacant and impossible lower half of the interval containing zero in any explicit fashion, and this leads to inconsistencies in the listing of class intervals. Perhaps the best way out is simply to write "0" in any list of class limits or class intervals, implying that the interval is a special case.

Construction of the frequency table. We have gone into the matter of class limits and class intervals more thoroughly than is necessary for actually drawing up a table. For with the true class limits in mind, we shall go right ahead and use the simpler designation of intervals in the table. Usually a tallying process is necessary after the intervals have been chosen, but if an array has been formed, one can simply count the cases falling within each interval from the array. Thus, it is a simple and easy transition from the array of Table 9 to the frequency table shown as Table 11.

As in all frequency tables, the left column specifies the class interval, and the right column designates the number of individuals whose measures fall within the specified intervals. Actually, what we have listed as

Table 11. FREQUENCY DISTRIBU-
TION OF 117 WHITE TENANT
FARM WOMEN BY NUMBER OF
CHILDREN BORNE, 1938

Number of children borne	Number of women
All numbers	117
0	1
1	8
2	10
3	5
4	14
5	10
6	16
7	14
8	10
9	6
10	6
11	7
12	5
13	4
14	0
15	0
16	1

Source: Table 8.

the specifications of the class intervals in Table 11 are the midpoints of the class intervals. This is done since the observed values can take only this value. For later analysis we shall consider these intervals to be as shown in Table 10.

Use of the frequency table. Now let us notice what advantages this frequency table has over the ungrouped data of Table 8 or the array of Table 9. In the first place, it is more concise than either of the others. This quality is valuable not only for saving space and thereby expense in a report, but also for enabling a reader to comprehend the data more readily. Thus, we see that putting data into a frequency table fulfills the *condensing* or *summarizing* function of descriptive statistics even though it does not reduce the information to *one* summarizing measure. Instead of having the numeral "6" as number of children borne appear in the frequency table 16 times, as it does in Tables 8 and 9, we now have the numeral "6" only once, with the numeral "16" indicating its frequency.

This leads to the second advantage, which is the efficient preparation of results for later analysis. As will be seen, instead of treating each value of each measure separately, we shall be able to treat all the measures of one class simultaneously and then multiply the result by the frequency of that class. Finally, if the array had not been formed and if one compares
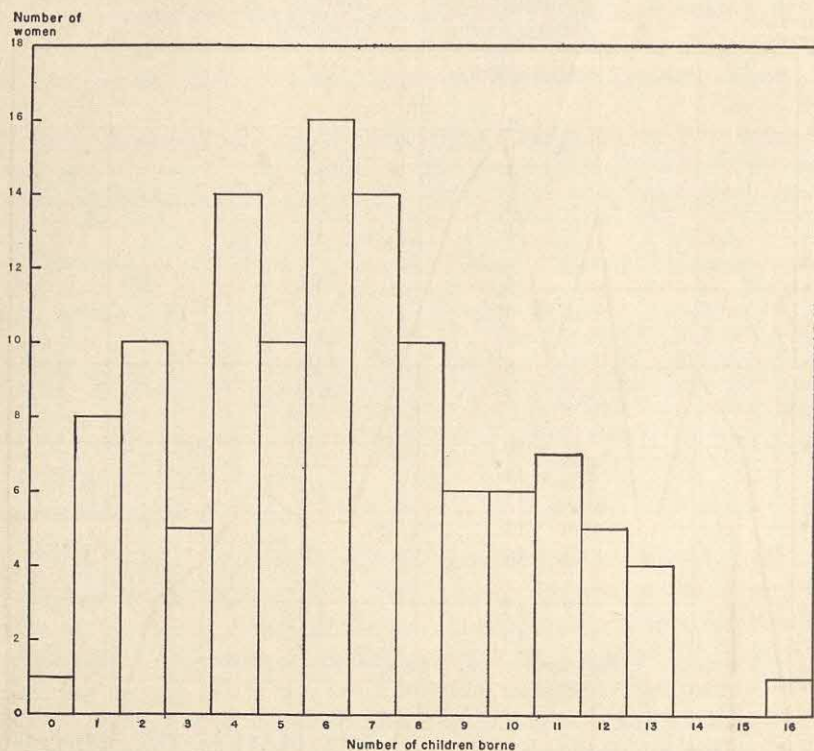
Figure 10. Frequency Distribution of 117 White Tenant Farm Women by Number of Children Borne, 1938. Histogram with 1-child intervals. (Source: Table 11.)

directly Tables 8 and 11, the advantage of the organization of the data into a frequency table for affording a better grasp of the distribution of fertility among these women as a whole is more evident. After we have investigated the aspects of central tendency, dispersion, and form, we shall see that the frequency table offers information on all these aspects although it does not afford precise, single, summarizing measures of them.

Graphic presentation of the frequency distribution. We have already noted that histograms and coordinate charts are both appropriate forms for presenting frequency distributions graphically. Figures 10 and 11 are

a histogram and a coordinate chart for the frequency distribution of Table 11. The differentiation between a histogram as an area diagram and a coordinate chart as representing points located with reference to two axes was explained in Chapter 5. This differentiation can be explained more fully now that the matter of class limits has been discussed.
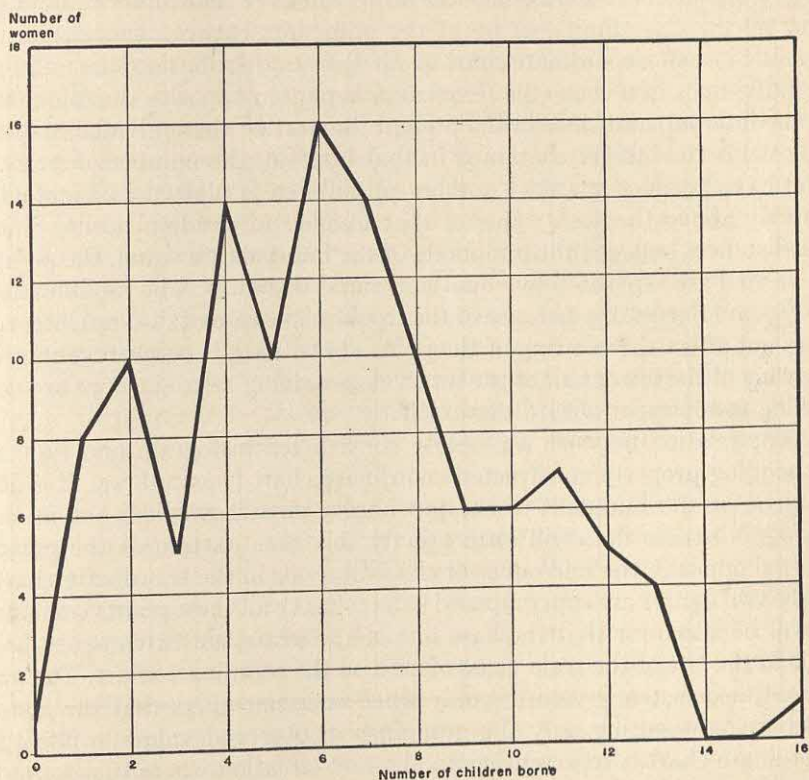
Figure 11.   Frequency Distribution of 117 White Tenant Farm Women by Number of Children Borne, 1938. Coordinate chart with 1-child intervals. (Source: Table 11.)

First, let us call attention to the fact that the base line of the histogram is not necessarily the scale of a continuous variable. The histogram is based on the groupings of observed measures indicated in Table 11 before any assumptions have been made as to continuity of the variable formed by measures on the characteristic, "number of children borne." Since each interval, including the zero interval, covers only one possible value of an observed measure, we say that the intervals are equal, although strictly speaking the intervals have no width—they are only points. The numbers, 0, 1, 2, 3, and so on, just below the base line in the

histogram are not scale values but are designations of the class groupings of the observed integral measures.

The transition from the histogram of Figure 10 to the coordinate chart of Figure 11 is the graphic equivalent of the process of assuming continuity of the variable formed by the measures. The observed frequencies in any one interval are assumed to be spread over that interval, defined now by the class limits set up at the midpoints between successive integers. Since in a coordinate chart of a frequency distribution, the number of individuals in a class (the frequency) is plotted opposite the midvalue of the interval, and since in the present illustration the midvalue of each interval is the integer contained in that interval, the number of women who have borne a specified number of children is plotted in Figure 11 directly above the scale value of that number of children borne. Since the distances between the midpoints of the intervals are equal, the points so located are separated by equal horizontal distances. The modification in Figure 11 necessary because of the troublesome case of the zero interval does not affect the plotting of the point above zero; it only prevents the showing of the lower half of the interval containing zero, since we are not willing to represent absurdities explicitly.

The relation between a properly constructed histogram and its corresponding properly constructed coordinate chart is as follows. If a dot is made at the midpoint of the top border of each vertical bar in the histogram, these dots will form exactly the same pattern as the points plotted opposite the midvalues of class intervals in the coordinate chart. If the two figures are superimposed so as to make all these points coincide, it will be apparent that the base line of the histogram extends one-half unit to the left of the scale value of zero on the coordinate chart. The apparent inconsistency vanishes only when one remembers that the histogram is representing only the groupings of observed values, while the coordinate chart is representing an abstract situation where the assumption of continuity has been made. It is to emphasize this differentiation that we follow the convention of specifying the actual groupings of observed measures as the designations of the different vertical bars in a histogram and of specifying scale values as the designation of the points on the horizontal axis in a coordinate chart.

Use of histogram or chart to determine proper class intervals. It will be noted that there is a zigzag appearance to the histogram and to the coordinate chart, which means that the frequencies change erratically from one class to the next. This erratic fluctuation can be detected in the frequency table, but it is more obvious from inspection of the graphic forms. Now the question of whether we should use this grouping of measures into class intervals or whether we should choose some broader class intervals which would even out some of the irregularities brings up

a fine point of differentiation between descriptive and inductive statistics. If we followed to its logical conclusion the differentiation set forth in the preceding chapters of this book, we should be forced to the position that, irregularities or no irregularities, Table 11 and Figures 10 and 11 describe the actual facts of the distribution of fertility in this group of 117 white tenant farm women in 1938, and since such a description is all we are after in descriptive statistics, the present grouping is satisfactory. However, the position one infers that most statisticians take (few writers differentiate so explicitly between the functions of descriptive and inductive statistics as we have done) is the following. While it is true that we are not actually generalizing any of the results of this analysis to a universe, yet *any* research of scientific value is aiming eventually at some sort of generalization, even if to an imaginary universe, and our ideas about fertility *in this particular group* will be clarified if we put the description of the distribution into a form more similar to that which would represent the distribution of fertility in a universe of an infinite number of women, where no irregularities in classes would show up due to the smallness of the number of cases studied. Moreover, for practical reasons, we try to perform our descriptive analysis in such a way that if the group described is a sample, we can later generalize the results to a universe.

A second grouping of measures into class intervals. Therefore, let us try grouping the measures into wider class intervals, intervals which include two adjacent numbers of children borne. We have the choice of grouping them in two ways, either thus, 0 and 1, 2 and 3, 4 and 5, and so on, or thus, 0, 1 and 2, 3 and 4, 5 and 6, and so on. The second grouping is preferable, since the interval including zero will either include negative values or will be of a size unequal to the others in either way of grouping, and the troublesome interval will be easier to treat if all the measures in it have the one value, zero. Another reason for choosing the second grouping is that it makes a smoother distribution, a desired objective. Table 12 is the frequency table resulting from the second grouping. As before, the simpler designations of class intervals are used in the table, but as before we must keep in mind the exact values of the class limits for later analysis of these frequency tables. The midvalues of intervals are now no longer integers. The precise specifications of the intervals after the first are 0.5 up to but not including 2.5, 2.5 up to but not including 4.5, and so on. The conventional specifications, which are less awkward, are 0.5–2.4, 2.5–4.4, and so on. The midvalues of the class intervals (after the zero interval) are 1.5, 3.5, and so on. We shall use these midvalues later when we perform analysis of grouped data on the assumption that all the measures contained in a class interval may be treated as if they had the midvalue of that class interval.

Table 12. FREQUENCY DISTRIBUTION AND PER-
CENTAGE DISTRIBUTION OF 117 WHITE TENANT
FARM WOMEN BY NUMBER OF CHILDREN BORNE,
1938

Number of children borne	Number of women	Percent
All numbers	117	100.0
0	1	.9
1 and 2	18	15.4
3 and 4	19	16.2
5 and 6	26	22.2
7 and 8	24	20.5
9 and 10	12	10.3
11 and 12	12	10.3
13 and 14	4	3.4
15 and 16	1	.9

Source: Table 8.
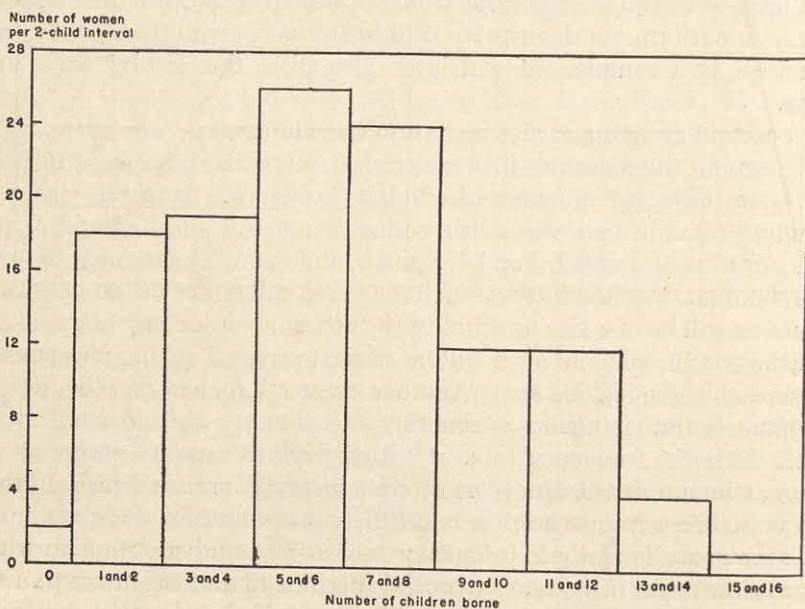
Figure 12. Frequency Distribution of 117 White Tenant Farm
Women by Number of Children Borne, 1938. Histogram with
2-child intervals. (Source: Table 12.)

**Graphic presentation of the frequency distribution resulting from the
second grouping.** Figures 12 and 13 have been constructed to present
graphically the frequency distribution of Table 12, just as Figures 10 and

11 present graphically the frequency distribution of Table 11. With the exception of the treatment of the interval containing zero, no new explanation is needed. As before, the vertical bars of the histogram are specified by the designations used in the grouping of the observed measures, 1 and 2, 3 and 4, and so on; and as before, the horizontal scale of the coordinate chart has scale designations of values of the variable, which has been assumed to be continuous. In treating the zero interval, the
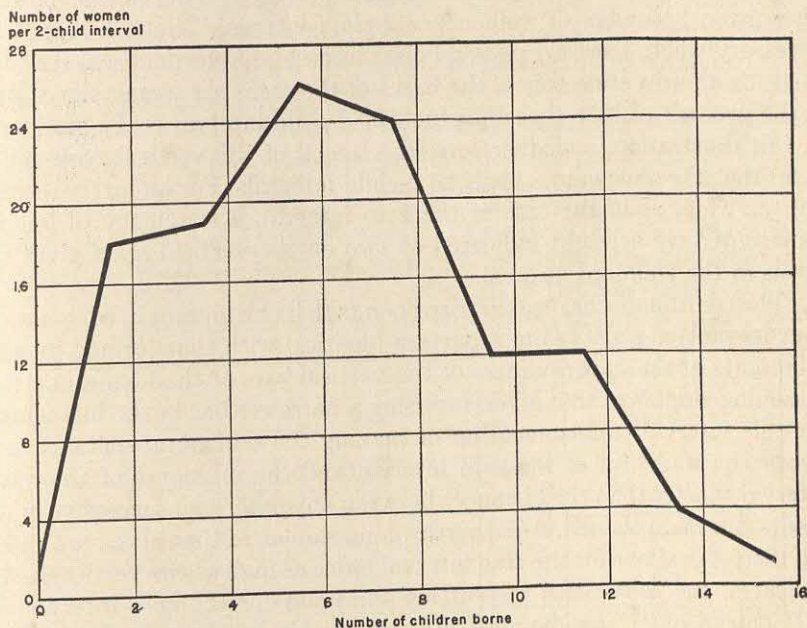
Figure 13. Frequency Distribution of 117 White Tenant Farm Women by Number of Children Borne, 1938. Coordinate chart with 2-child intervals. (Source: Table 12.)

same cutting off of the absurd negative lower half of the interval in the coordinate chart has been done as before.

Conventions for graphic presentation of a frequency distribution with unequal intervals. The zero interval in Figures 12 and 13, however, has still another irregularity to cause trouble. Even in the grouping specified on the histogram, the interval containing zero cannot be considered to be equal to the other intervals, for it includes only one value which observed measures can take, while all the other intervals now each contain two values which observed measures can take. It is conventional in such a case to represent an interval which is only half as "wide" as the others by a vertical bar only half as wide as the other bars in the histogram. This

seems reasonable enough, but the convention which is a little more diffi-
cult to explain is that the half width bar is made twice as high as its
frequency indicates on the vertical scale. This is done because a histogram
is an *area* chart, and the one woman in the zero interval must be repre-
sented by exactly the same area as a woman in any other interval. If the
base of her bar is only half as wide as those of the other bars, the height
must be twice as great as that required for the other bars to represent
one woman. Note the peculiar legend over the vertical scale of the
histogram, "Number of women per 2-child interval." Such a legend is
necessary where there are unequal class intervals. A vertical scale can be
used for an area scale only if the bases of the areas are equal, since area
is the product of base times height. The 2-child interval is the standard
in this illustration, and, therefore, the legend of the vertical scale indi-
cates that the scale values apply to 2-child intervals. For an interval only
half as wide, as in the case of the zero interval, a frequency of one is
represented by a height indicated as two on the vertical scale given in
terms of the standard interval.

The coordinate chart again corresponds to its histogram in such a way
that its plotted points form a pattern identical with that formed by the
midpoints of the upper borders of the vertical bars of the histogram. By
reasoning similar to that given for using a narrower bar in the histogram
for this interval, the convention of making the horizontal distance be-
tween the midpoint of the zero interval and the midpoint of the next
interval shorter than the distances between any other two successive mid-
points can be explained. And by reasoning similar to that given for mak-
ing the vertical bar for the zero interval twice as high as the vertical scale
indicates, the convention of plotting the point for the zero interval at
"2" instead of "1" on the vertical scale can be explained. Another sug-
gestion may help the student to see why it is "right" to plot the frequency
of the zero interval at "2" instead of at "1." Although the classes are
different, the two coordinate charts shown as Figures 11 and 13 do repre-
sent the same observed distribution, and the patterns formed by lines
joining their plotted points should be approximately similar, even though
the one with broader intervals should have fewer irregularities. In Figure
13 the ratio of the height of the point plotted for the zero interval to the
height of the point plotted for the next interval should be the same as the
ratio in Figure 11 of the height of the point plotted for the zero interval
to the average of the heights plotted for the next two intervals (which
have been combined to form one interval in Figure 13). The ratio will
be the same, and the pattern of the two descriptions of the same distri-
bution will be similar, *only* if the point for the zero interval is plotted at
"2."

The percentage distribution. In addition to the columns indicating class intervals and class frequencies, Table 12 has a third column which shows the percentage each class frequency is of the total frequency. Such a percentage distribution is especially useful when distributions for groups of different sizes are being compared. For instance, in a similarly grouped percentage distribution of 1,000 urban women by number of children, we could read off what percentage of the urban women had borne five or six children and compare it with the 22.2 percent of the tenant women who have borne that number. The percentages of the third column are obtained by the methods explained in the preceding chapter for computing component percentages.

The cumulative frequency table. It is possible that we may wish to know how many of the women have borne a given number of children or

Table 13. CUMULATIVE FREQUENCY DISTRIBUTION AND CUMULATIVE PERCENTAGE DISTRIBUTION OF 117 WHITE TENANT FARM WOMEN BY NUMBER OF CHILDREN BORNE, SHOWING THE NUMBER OF WOMEN HAVING BORNE FEWER THAN THE STATED NUMBER OF CHILDREN, 1938

Number of children borne	Number of women	Percent
Fewer than 0.........	0	.0
Fewer than 1.........	1	0.9
Fewer than 3.........	19	16.2
Fewer than 5.........	38	32.5
Fewer than 7.........	64	54.7
Fewer than 9.........	88	75.2
Fewer than 11.........	100	85.5
Fewer than 13.........	112	95.7
Fewer than 15.........	116	99.1
Fewer than 17.........	117	100.0

Source: Table 12.

more, or how many have borne fewer than a given number of children. Such information can be obtained by adding the appropriate frequencies of Table 11 or 12. A modification of the frequency table with all these additions performed is called a cumulative frequency table. From Table 12 we obtain the cumulative frequencies of Table 13 by beginning with zero, as obviously the number of women who have borne fewer than zero children. To obtain successive entries in the cumulative frequency column, we add the frequencies and take a subtotal as the cumulative frequency after each addition. The resulting number is the number who

have borne fewer children than the smallest number in the next interval as is indicated in Table 13. If we perform the same operations beginning at the other end of the distribution, we get an "or more" cumulative frequency table.

The cumulative percentage distribution. The last column in Table 13 shows what percentage each entry in the middle column is of the total
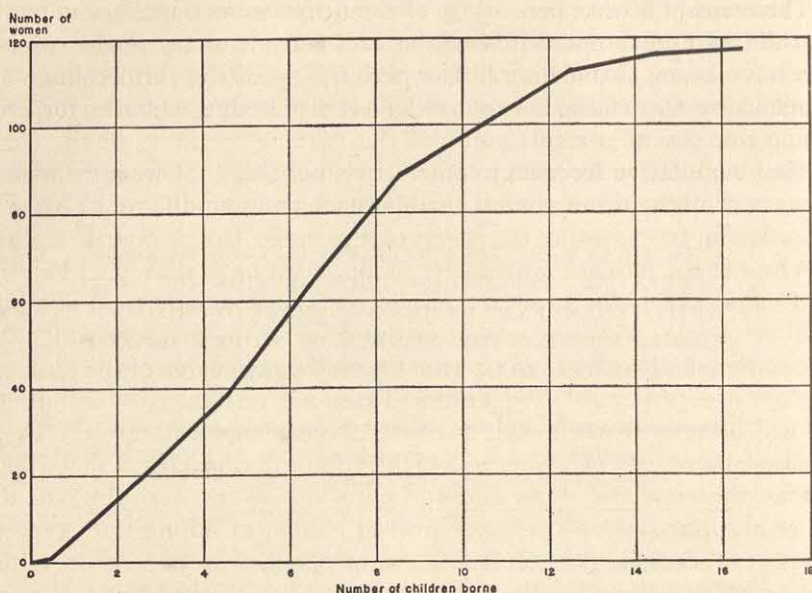
Figure 14. Cumulative Frequency Distribution of 117 White Tenant Farm Women by Number of Children Borne, Showing Number of Women Having Borne Fewer than the Stated Number of Children, 1938. Coordinate chart with 2-child intervals. (Source: Table 13.)

number of women. Such percentages are especially useful in making comparisons with other groups of different sizes.

Graphic representation of cumulative distributions. The graphic representation of a cumulative distribution expressed either as a frequency distribution or as a percentage distribution is called an ogive. Figure 14 shows the "fewer than" ogive corresponding to Table 13. The coordinate chart is based on the assumption of continuity of the variable, "number of children borne." Therefore, the points are plotted not above the integers but above the lower limit of the class containing the integer. This is necessary because the assumptions of continuity and of grouping mean that we must consider the measures of one and two children as extending all the way down to 0.5 child. These "fewer than" and "more than"

cumulative distributions are useful for emphasizing different features of the distribution, while describing the distribution as a whole.

DESCRIPTION OF THE CENTRAL TENDENCY
OF THE DISTRIBUTION

The concept of average. With the description of the frequency distribution as a whole accomplished, we are ready to go on to the specific aspect of a quantitative distribution which is usually described first—its "central tendency." The concept of an average is an abstraction created to help one choose a single value of the varying measures of the characteristics which can be used to represent all of them. There are different averages or measures of central tendency which have different points of advantage in representing the group of measures. The *arithmetic mean* is the value of the measure which every unit would have if they all had the same value and if the total amount of the characteristic for the group remained as observed; the *median* is the value of the measure which the middle item of an arrayed group has; the *mode* is the value of the measure which is observed most frequently. (These are introductory definitions and will be refined when each measure is considered separately.) There are also other sorts of averages which require mathematical definitions such as the *geometric mean* and the *harmonic mean*. According to the nature and purpose of a research project, the most appropriate type of average for describing the central tendency of a distribution varies. Often we may wish to use all of the first three named to show variation among the averages. But no matter which one we use, it is difficult to conceive of a description of the incidence or distribution of a measurable characteristic which does not need to employ one of these condensing measures.

The use of averages reflects the tendency in the thinking of the layman to try to pick a representative unit of varying ones to characterize a class. Such phrases as "the average man on the street" or "the typical adolescent" are examples of this tendency. The major point of difference between the statistical and the lay problem of getting representative items or units is that in the statistical problem the aim is to secure representativeness in only one or in a limited number of specified measurable characteristics; while in the nonstatistical problem, the representativeness sought is usually for an unlimited number of characteristics, most of them not at present measurable. Hence, the solution for the statistical problem can be far more exact, precise, and verifiable than that for the nonstatistical.

The arithmetic mean. The most important and most useful measure of central tendency is the *arithmetic mean*. It is this summarizing measure

that is usually referred to when one speaks of an "average" and it is this particular sort of mean that is referred to when one uses the word "mean" without a qualifying adjective. Most of the readers have probably added up a series of values and divided by the number of values added to get an average. This is exactly what we shall do in statistics, but we shall formulate the process in symbols, and shall also develop short methods of performing the process for grouped data.

Computation of the mean from ungrouped data. If we add the numbers of children listed in Table 8, we get 742 as the sum; if we divide 742 by 117, we get 6.34 as the arithmetic mean of the number of children borne by the women of this group.[2] Let us note what has been accomplished in the way of condensation by this process. Instead of 117 records of number of children borne, we now have one figure, 6.34, which is based upon all 117 observations and which can be used to represent all of the 117 observations as an average of them. Of course, it does not tell us anything about how the individual women vary about the average; a measure of dispersion is required to do this. Nor does it tell us that one woman had 16 children while another had none, for detail is almost always sacrificed in the process of condensation. But it does give a summarizing measure of the fertility level of the group, which all of the 117 observations considered singly cannot afford and which can be used to compare this group with others in regard to fertility. If any one single figure must be chosen to describe the incidence of fertility among this group, this arithmetic mean is probably the best choice, for it tells more than any other single figure about the level of incidence of the characteristic being studied.

Since arithmetic means are so indispensable for condensing information and since almost all further analysis is based upon them, we shall express the procedure for obtaining an arithmetic mean in terms of a formula. In this formula we shall begin to introduce the notation which will be used throughout this text.

When considering any single distribution of a characteristic among a group of units, we shall denote the measures of the characteristic for the units, in the order in which they are measured or listed as,

$$X_1, X_2, X_3, X_4, \ldots, X_N$$

where the three dots, $\ldots$, are to be read "and so on up to," and where N is equal to the number of observations or measures, in the case of our

---

[2] A conventional rule of thumb is to carry computed measures, such as the mean and standard deviation, one decimal place further than the measures on which they are based. Since in this chapter we shall be interested in comparing computed measures which are very close together and since we shall compute further measures from the computed ones, we are carrying the results to two decimal places rather than one. It is advisable to carry results to extra decimal places whenever further computations are going to be based upon them, although the published results should show only the places justified.

illustration, 117. Now if we let $\bar{X}$ be the symbol for the arithmetic mean of the measures, we can write the formula for $\bar{X}$, thus,

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} \tag{1}$$

This algebraic formula translated into words says, "The arithmetic mean is equal to the sum of the value of the first measure, plus the value of the second measure, plus the value of the third measure, and so on up through the value of the last measure, divided by the number of measures." To indicate the process of adding the items of a series, we have a symbol, Σ, which is read "the sum of all of the" or "the summation of the." Using this summation symbol enables us to write formula (1) in a shortened form, thus,

$$\bar{X} = \frac{\Sigma X}{N} \tag{2}$$

When the summation sign, Σ, is used as above, without any limits of summation expressed, it is understood to mean that every one of the X's is to be included in the summation.

Computation of the mean from grouped data. Formulas (1) and (2) are alternate ways of defining the procedure for obtaining the arithmetic mean from ungrouped data. For our example this involves adding 117 figures, which is not a lengthy matter if an adding machine is available, but for examples involving great numbers of cases, this method takes too long to be practicable. Let us consider by way of illustration of the transition from ungrouped to grouped data, the computation of the arithmetic mean from Table 11. Remembering that what we want for the numerator of formula (1) is the total number of children borne by the 117 women, we see that instead of adding each woman's number of children borne separately, we can multiply any number of children borne, such as 6, by the number of women who have borne this number, 16. By this one multiplication we save 16 additions. We can compute the mean from Table 15 by a modification of formula (1), thus,

$$\bar{X} = \frac{1(0) + 8(1) + 10(2) + 5(3) + 14(4) + \cdots + 1(16)}{117}$$

$$= \frac{742}{117} = 6.34$$

We have used 17 multiplications and 17 additions, a total of 34 operations in the place of the 117 additions called for by formulas (1) or (2). What we have done is to multiply the frequency of a class interval by the mid-value of that interval and then to add these products. If we let m repre-

sent the midvalue of an interval and f the frequency of that interval, we can express the procedure we have just used for finding the mean from grouped data in a formula, thus,

$$\bar{X} = \frac{\Sigma fm}{N} \qquad (3)$$

Note that the value obtained for the arithmetic mean, 6.34, is exactly the same when computed from data grouped by 1-child intervals as when computed from ungrouped data. This is true because in the process of grouping for Table 11, it will be remembered that we did not sacrifice any of the accuracy of the observations. The grouping assumption, that all the measures in a class can be treated as if their values are the midvalue of that class, is really not an assumption in this particular example, because here the value of every measure in each class is *precisely* the midvalue of that class, the only value it can take. Since this is not usually the case, however, it will not generally be true that a mean secured from grouped data will have identically the same value as the one secured from ungrouped data.

Let us now compute the mean from the grouped data of Table 12. Because of the broader groupings we shall have a smaller number of multiplications, but they will involve decimals, since the midvalues of these class intervals are not integers. In all dealings with the data grouped as in Table 12, we shall have to consider the interval containing zero as a special case, which will make the work a little more complicated, but will serve to illustrate further the sorts of modifications in procedures and formulas necessitated by the use of unequal intervals, which sometimes cannot be avoided.

Table 14 shows the computations which are necessary in order to compute the mean from the grouped data of Table 12. Column (1) gives the class limits, and column (2) gives the midpoints. The frequencies shown in column (3) are taken directly from Table 12. The midpoint of each class interval is obtained by taking the mean of its true class limits (see Table 10), not of its specified limits. Notice also that zero is written as the class limits of the lowest interval; this is done to avoid using the value -0.5, which would be necessary if we continued our assumption of continuity into the negative zone. This treatment is also consistent with the convention of considering as midpoint the actual mean value of measures observed in unequal intervals.

The entries in column (4), designated as fm, are obtained by multiplying together the corresponding entries in columns (2) and (3). For example, the entry in the second row, 27.0, is obtained by multiplying together 1.5 and 18.

We sum the entries in column (4) to obtain Σfm, shown in Table 14 as 738.0. We can now evaluate formula (3),

Table 14. COMPUTATIONS FOR MEAN FROM GROUPED DATA OF TABLE 12

Class limits (1)	m (2)	f (3)	fm (4)	d' (5)	fd' (6)	F (7)
0	0.0	1	0.0	−2.75	−2.75	1
0.5–2.4	1.5	18	27.0	−2	−36	19
2.5–4.4	3.5	19	66.5	−1	−19	38
4.5–6.4	5.5	26	143.0	0	0	64
6.5–8.4	7.5	24	180.0	1	24	88
8.5–10.4	9.5	12	114.0	2	24	100
10.5–12.4	11.5	12	138.0	3	36	112
12.5–14.4	13.5	4	54.0	4	16	116
14.5–16.4	15.5	1	15.5	5	5	117
Sums		117	738.0		47.25	

$$\bar{X} = \frac{\Sigma fm}{N} = \frac{738.0}{117} = 6.31$$

$$\bar{X} = \bar{X}' + \frac{\Sigma fd'}{N} \, i = 5.5 + \frac{47.25}{117} \, (2) = 6.31$$

$$\bar{X} = \frac{\Sigma fm}{N} = \frac{738.0}{117} = 6.31$$

Notice that this value differs slightly from the previous value obtained, 6.34. The difference is not due to mistakes in computation, but to the divergence of the data from the assumption made that the midpoint of a class is the arithmetic mean of all the values in the class. Which is the "right" or "correct" value of the mean? As far as the group being described is concerned, the value from the ungrouped data is the correct value of their arithmetic mean, although the value obtained from grouped data is usually so close to the correct value that its use is permitted to save time.

Computation of the arithmetic mean from grouped data by the short method. This short method on first inspection appears roundabout, for the procedure is to guess the value of the mean, then to figure how far we have missed our guess, add this amount to the guessed mean, and thus obtain the correct mean. (The "correct" mean will have the same value as the one obtained by the long method with grouped data, not the same as the one obtained with ungrouped data.)

This short method of obtaining the mean is based upon a very important property of the mean—that the algebraic sum of the deviations

of the individual measures from the mean is zero. For example, the mean of the numbers 3, 2, 6, and 9 is 5. The deviations from the mean are -2, -3, 1, and 4 respectively. The sum of these deviations is zero. We use the values of the deviations of the individual measures from the mean so frequently that a special symbol, x, is assigned to the deviation from the mean. We attach subscripts to these small x's when necessary for distinguishing them, just as we do to the large X's. Obviously x_1 refers to the deviation of X_1 from the mean. For the general case, x is defined by the equation,

$$x = X - \bar{X} \tag{4}$$

Now the property of the mean which we have been discussing above may be stated algebraically, thus,

$$\Sigma x = 0 \tag{5}$$

Because of this property we can determine how much we have missed the correct mean by our guessed mean. We compute the deviations from our guessed mean and add them. If the sum of these deviations happens to be zero, this means we have guessed the mean correctly; if the sum is positive, this means we have guessed the mean too low; and if the sum is negative, this means we have guessed the mean too high. When the sum is different from zero, either positive or negative, we can determine how far we have missed the mean by dividing the sum by the number of measures whose deviations we have added. This tells the distance we would have to shift the guessed mean to make the sum of the deviations add up to be zero, and therefore locates the true mean.

In practice when we use this method with grouped data, we always guess the mean to be at the midvalue of one of the class intervals. Instead of computing the deviations from the guessed mean in terms of the original units of measurement, we compute them in terms of "step deviations," where a "step" is the size of a class interval.

The "step deviation" of any class is symbolized by d' and is the distance from the guessed mean to the midpoint of the class measured in "step" units. It can be symbolized as

$$d' = \frac{m - \bar{X}'}{i} \tag{6}$$

where $\bar{X}'$ is the guessed mean
m is the midpoint of the class
and i is the size of the usual class interval in the distribution.
The device of using step intervals provides us with small whole numbers for multiplying and reduces the length of time required for computation.

Returning to our problem in Table 14, in this example we know within which interval the mean falls. However, the method works just as well if we do not know and guess the wrong interval. Since we want

the guessed mean, $\bar{X}'$, to be the midpoint of an interval, let us guess it to be 5.5. In column (5) of Table 14 we have filled in the step deviation, d', of each class using formula (6). When all the classes in a distribution are of equal width, the d''s can be written down by inspection. However, for an unusual interval such as our zero interval we resort to formula (6), remembering that i, the usual class interval, is a constant for any given distribution. Computing d' for our zero interval we have

$$d' = \frac{m - \bar{X}'}{i} = \frac{0 - 5.5}{2} = 2.75$$

Column (6) in Table 14, labeled fd', is the product of the frequency of each interval (column 3) times the step deviation of that interval (column 5). The summation of the fd' column gives $\Sigma fd'$, though we must take into account the signs of the fd''s in summing. To find how much we have missed the true mean by our guess, we divide this sum, 47.25, by the number of deviations which were added, 117, and get .404. It must be kept in mind that we are performing these computations in terms of class intervals and to transform the correction just found into number of children borne, we must multiply it by the class interval, 2, getting 2(0.404) or 0.808 children. Since fd' was positive we must add this to our guessed mean, thus,

$$5.5 + 0.808 = 6.308 \text{ or } 6.31$$

This procedure for obtaining the mean by the short method with grouped data can be expressed by a formula, thus,

$$\bar{X} = \bar{X}' + \frac{\Sigma fd'}{N} i \qquad (7)$$

where $\bar{X}$ = mean
  $\bar{X}'$ = guessed mean
  f = frequency in a class interval
  d' = step deviation of a class interval
  N = number of measures
  i = width of usual class interval in the distribution.

The value of the mean obtained by this method is exactly equal to that obtained by the longer method with grouped data.

The median. The mean is not always the most appropriate measure of central tendency, although on the whole it is the most useful. Perhaps the second most important measure of the central tendency of the distribution of a variable characteristic is the median. The *median* is sometimes roughly defined as the middle measure in a series. It is more precisely defined as that value of the measure which divides a series of measures into two equal groups, one group with measures of higher value

and one group with measures of lower value. Neither of these definitions is completely unambiguous, as we shall see from illustrations. It will be necessary to define the median operationally by formulating the processes of computing it for ungrouped and grouped data.

Computation of the median from ungrouped data. Since certain complications are involved in the computation of the median number of children borne by the women in our example, let us consider first a smaller group of five women who have borne the following numbers of children: 4, 5, 6, 7, and 8. By either of the above definitions, the median number of children borne by these five women is 6; 6 is the middle measure in the series, and it is also the value of the measure which divides the series of measures into two equal groups (if we can imagine the measure 6 itself split in two parts so that half of it can go with either group). If we consider only the first four women as the group, however, with their numbers of children, 4, 5, 6, and 7, we cannot use the first definition because there is no middle measure in an even number of observed measures; and if we use the second definition, we can call any value between 5.0 and 6.0 the median. It is the customary practice in such a case to take the value midway between 5.0 and 6.0—or their arithmetic mean—5.5 as the median, although this is not usually explicit in the verbal definition of the median.

An attempt to apply either of the two definitions to the example of the 117 tenant farm women brings up additional difficulties. Since the number of women is an odd number, 117, it seems that there should be a middle measure which would be the measure of the 59th woman when they are ranked in order of their number of children. If each woman had a different number of children, this would be a feasible procedure, but the measures of all 117 women, it will be remembered, take only 15 different values. In the array of Table 9 there was no woman who ranked as 59th; the 14 women who had 7 children all had the rank of 46.5, and the 16 women who had 6 children all had the rank of 61.5. A case could be made for calling 6 the median number of children; since the rank 61.5 was determined by averaging all the ranks from 54 through 69 to get the rank 61.5, one could say that the group who have borne 6 children includes the woman with the middle measure. On the other hand, it is not at all clear how the value 6 divides the measures into two groups; there are 48 measures less than 6 and 53 greater than 6, while there are 16 measures with a value of 6. A way out of this situation, where the ungrouped data are actually "grouped" from being bunched, is to take an intermediate value between the value 7, which has a rank of 46.5 and the value 6, which has a rank of 61.5. The process is one of linear interpolation, which means finding an intermediate value between two known values by assuming that the value increases evenly throughout the interval. We know that

the 46.5 rank has a value of 7 and that the 61.5th rank has a value of 6; the question is, What value would the 59th rank have? The difference between the two ranks with known values is

$$61.5 - 46.5 = 15$$

and the difference between the unknown and the higher of the known ranks is

$$61.5 - 59 = 2.5$$

Now if a change of 15 ranks corresponds to a change of 1 in value, we can easily find what value a change of 2.5 ranks corresponds to by simple proportion, thus,

$$x : 2.5 = 1 : 15$$

$$x = \frac{2.5 \times 1}{15} = 0.167$$

and the required value of the measure for the 59th rank is

$$6.0 + 0.167 = 6.167 \text{ or } 6.17$$

This is the best interpretation of the definition of the median for such problems, although it is not precisely the same value that would be obtained by finding the median of the distribution in Table 11 by the method of grouped data.[3]

Computation of the median from grouped data. We have anticipated some of the procedures used in computing the median from grouped data. When we turn to grouped data, we no longer seek the "middle" measure referred to in the first definition, but seek the position on the quantitative scale of measures which will divide the group of observed measures into two equal parts. Therefore, the first step is to divide the number of cases by 2, $\frac{117}{2} = 58.5$. This procedure seems to be slightly inconsistent with the one above which sought the 59th rank, but the median is defined slightly differently for grouped and ungrouped data. Of 117 ranks, the $\frac{N+1}{2}$ th or 59th is the middle one required by the first definition; but of 117 measures grouped into intervals, where the assumption of continuity has transformed points into intervals the $\frac{N}{2}$ th or 58.5th measure (an abstraction, of course) divides the group of measures into two equal parts.

---

[3] The value of the median obtained by interpolating between ranks is different from that obtained by interpolating between the limits of one class interval, as would be done in computing the median from the grouped data of Table 11, because in the first interpolation the number of measures in an adjacent interval affects the median, while in the second interpolation, the number of measures in the adjacent interval does not affect the value of the median. The method for grouped data is preferable.

In column (7) of Table 14 is shown the cumulative frequency up through each class. From this column we see that the 58.5th measure falls in the interval with class limits from 4.5 to 6.4, which contains all the cumulative frequencies from 38 to 64. We interpolate, assuming that the 26 frequencies in this interval are evenly distributed between the true limits of the interval, $4.5 - 6.5$. We wish to determine what distance from the lower limit of the interval the value occurs which corresponds to rank 58.5. Since $58.5 - 38 = 20.5$, we take $\dfrac{20.5}{26}$ of the interval width, 2, and add this to the lower limit of the interval, thus,

$$\frac{20.5}{26}(2) = 1.577$$

$$\text{Median} = 4.5 + 1.577 = 6.077$$

This process can be symbolized in a formula, thus,

$$Md = l + \frac{\dfrac{N}{2} - F}{f} i \tag{8}$$

where Md = median

N = number of measures

l = lower limit of the interval containing the $\dfrac{N}{2}$ th measure

F = cumulative frequency up to the interval containing the $\dfrac{N}{2}$ th measure

f = frequency of the interval containing the $\dfrac{N}{2}$ th measure

i = size of the interval containing the $\dfrac{N}{2}$ th measure

For grouped data it is best to regard formula (8) as the definition of the median since there is no way of stating the procedure precisely in a few words.

The mode. The mode is sometimes defined as that value of a measure observed most often in a series. As a measure of central tendency, it represents the most common value—the typical value. While the concept of the mode is easy to define roughly, the procedure for obtaining it is difficult to state precisely, since as in the case of the median, the procedures vary in different situations. The definition of the mode which is most acceptable from theoretical considerations involves a fitted curve, which will be discussed in Chapter 14.

Computation of the mode from ungrouped data. With ungrouped data, if at least for one value a considerable number of individuals have

the same measure, the value which is most frequently observed may be defined as the mode. In our example of 117 white tenant farm women it can be seen from the array of Table 9 that the mode, or modal number of children borne, is 6 because 16 women have this measure and not more than 14 have any other measure. If, however, there is no concentration of measures at one value, there is no way to compute the mode from ungrouped data.

Computation of the mode from grouped data. The first step in computing the mode from grouped data is to determine from inspection of a frequency table the modal interval, that is, the interval with the greatest frequency. In the frequency distribution of Table 12 it is easy to see that the modal interval is the one with class limits from 4.5 to 6.4, which has a frequency of 26. When the number of individuals studied is not very large, as in this example, one often stops with the determination of the modal interval and does not compute an exact value for the mode. We shall illustrate, however, one of the several procedures for specifying a particular value within the modal interval as the mode. This procedure is defined by the formula,

$$Mo = l + \frac{\Delta s}{\Delta s + \Delta g} i \qquad (9)$$

where l = the lower limit of the modal class

Δs = the difference between the frequency of the modal class and that of the class containing the next smaller values

Δg = the difference between the frequency of the modal class and that of the class containing the next greater values

i = the size (width) of the modal interval

Evaluating this formula with the data from Table 12, we have

$$Mo = 4.5 + \frac{(26 - 19)}{(26 - 19) + (26 - 24)} \times 2 = 6.055$$

The above procedure for obtaining a mode should be used only when from inspection of a graphic form of the distribution it is evident that there is one major point of concentration of values in the distribution. It is not advised where there appear to be two approximately equal points of concentration, a phenomenon termed "bimodality" of a distribution.

Other measures of central tendency. There are measures of central tendency other than the arithmetic mean, the median, and the mode, but they are not appropriate for the description of the distribution we are using as an example here. In fact, they are not at all widely used in the statistical analysis of data in sociological research. We give their definitions only by formula and refer the reader to the text in economics statistics listed at the end of the next chapter for further information about their computation and their use.

The *geometric mean* is sometimes used in problems relating to population growth and will be referred to in defining the coefficient of correlation. It is defined thus,

$$\text{Geometric mean} = \sqrt[N]{X_1 \cdot X_2 \cdot X_3 \cdots X_N} \qquad (10)$$

The computation of the geometric mean is usually performed by logarithms, in which case the following formula is used,

$$\text{Logarithm of geometric mean} = \frac{\Sigma \log X}{N} \qquad (11)$$

The *harmonic mean* is rarely used except for certain types of problems involving prices. It is defined thus,

$$\text{Harmonic mean} = \frac{N}{\Sigma \dfrac{1}{X}} \qquad (12)$$

There are still other measures of central tendency, but they are not generally applicable to the problems of sociological research.

SUGGESTED READINGS

See Suggested Readings at end of Chapter 9.
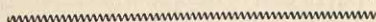
Quantitative Distributions: Measures of Dispersion and Form

THE concept of *dispersion* is less familiar than the concept of average. Yet, it will be recalled that the *variation* of individuals in a measurable characteristic is a basic condition for statistical analysis, and the concept of *variation* is fundamental to all statistical theory. If uniformity prevailed throughout the universe, there would be no need for statistical description, since any individual could be taken as a representative of all. On the other hand, if the distribution of characteristics were utterly chaotic and without regularity of form, statistical description of them would have little meaning although it would be mathematically possible. But actual observation has shown that many quantitative characteristics are so distributed that their measures for individuals cluster around some central or average value, while they depart from it in varying amounts. We have been concerned in the section on central tendency with determining and describing the central value; we now turn to the problems of determining and describing the departures of the measures from the central value, which aspect we call the spread, scatter, or dispersion of the distribution.

The range. The simplest measure describing the dispersion of a distribution is the distance on the scale of the values of the measures over which the observed individual measures are spread. The *range* is computed by subtracting the lowest observed value from the highest observed value, if the data are ungrouped. In our example, the range is

$$16 - 0 = 16$$

The range may be described in several ways—by its limits, by its extent, or by the extent of its two parts, the part below the mean and the part above the mean. Thus, we could say in describing the range for our example, "The range in number of children borne by these 117 women is from 0 to 16"; or "The range in number of children borne by the 117

women is 16"; or "The range in number of children borne by these 117 women extends 6.3 below the mean of 6.3 children borne, and 9.7 above the mean." For this particular example the first statement is perhaps preferable, but for other cases either of the two latter types of statements might be better.

If the data are grouped, the upper limit of the range is taken as the midvalue of the highest interval in which there are frequencies, and the lower limit of the range is taken as the midvalue of the lowest interval in which there are frequencies. Thus, in our example the grouping of data has the effect of diminishing the range by one half a child, for the data of Table 12 show as a range

$$15.5 - 0 = 15.5$$

The chief limitation of the range in describing the dispersion of the measures observed is that it is based upon only two observations, the highest and the lowest, and hence offers no information at all as to how the other measures are scattered, except that they are within these bounds. Yet as a first step in describing the dispersion of a distribution, it is useful because it does give the limits of variation and because it is so quickly computed and so readily understood.

The mean deviation. A measure of dispersion easily comprehended is the *mean deviation*, often known as the *average deviation*, which is defined as the arithmetic mean of the amounts by which the measures differ from the central value. We have noted as a property of the mean that the algebraic sum of all the deviations from it is zero. Therefore, in computing the average deviation from the mean we take the sum of the *absolute values* (signs disregarded) of all the deviations from the mean, and divide this sum by the number of deviations. The procedure with ungrouped data is tedious, for it involves subtracting the value of the mean from each measure, adding the absolute values of these differences, and dividing the sum by the number of measures. With the introduction of the symbol, $|\ |$, for "absolute value of," and the symbol MD for "mean deviation from the mean," we can state this procedure in a formula, thus,

$$MD = \frac{\Sigma\,|\,X - \bar{X}\,|}{N} \tag{1}$$

and since

$$X - \bar{X} = x$$

$$MD = \frac{\Sigma\,|\,x\,|}{N} \tag{2}$$

The equivalent formula for grouped data is

$$MD = \frac{\Sigma\,|\,fd\,|}{N} \tag{3}$$

where f = frequency of a class interval

d = deviation of class midpoint from the mean = $m - \bar{X}$

N = number of cases

The mean deviation from the mean is so little used in modern statistics that we shall not illustrate its computation.[1]

Notice that in referring to the mean deviation we have specified, "from the mean." We did this to differentiate between the "mean deviation from the mean" and "the mean deviation from the median," a measure which is sometimes used. The mean deviation from the median will always be smaller than the mean deviation from the mean (if the mean and median are not identical), since a property of the median is that it is the point from which the sum of the absolute values of the deviations of the measures will be a minimum. The procedures used in computing the mean (or average) deviation from the median are exactly the same as those described for computing the mean deviation from the mean, except that the deviation is computed from the median instead of from the mean in the case of either grouped or ungrouped data.

The quartile deviation or the semi-interquartile range. Before defining this measure of dispersion, we shall have to define certain measures of position which considered singly are measures neither of central tendency nor of dispersion, yet are analogous to the median and are used together to give a measure of dispersion. Just as the median is defined as the value of the variable which divides the distribution of measures into two parts with equal numbers in each, so the *quartiles* are defined as the three values which divide the distribution of measures into four parts with equal numbers in each. There is considerable confusion in the literature over two matters in connection with quartiles: the first, over the order in which the quartiles are numbered; the second, over the dual meaning of "quartiles" to refer both to the positions which divide a distribution into four parts and to the parts themselves. Let us clarify these points. If we consider the three quartiles, Q_1, Q_2, Q_3 as measures of position, Q_1 will fall nearest the lower end of the scale of measures (where the measures have the lowest numerical values); Q_2 will be identical with the median; and Q_3 will fall nearest the higher end of the scale of measures. Now if we are considering the three quartiles as names for the four parts cut by the three positions described above, we shall call the "first quartile" that part of the distribution between the lower limit of the range and Q_1, the "second quartile" that part between Q_1 and Q_2 (the median), the "third quartile" that part between Q_2 and Q_3, and the "fourth quartile" that part between Q_3 and the higher limit of the range. Note that the 25 percent of the individuals whose measures are *highest* are said to be in the

---

[1] For a short method of computing the mean deviation from the mean, see Robert Emmett Chaddock, *Principles and Methods of Statistics* (Boston: Houghton, 1925), pp. 156–158.

fourth quartile. The fourth quartile is sometimes referred to as the "upper quartile" or the "top quartile" or the "highest quartile." Finally, let us call attention to the slightly confusing fact that the individuals in the *fourth* quartile have the *highest* measures and the *lowest* ranks (in numerical values).

The computations for quartiles are very similar to those for the median; hence, we shall illustrate them only for grouped data. The formula for the first quartile is

$$Q_1 = l + \frac{\frac{N}{4} - F}{f} i \qquad (4)$$

where N = number of measures

l = lower limit of the interval containing the $\frac{N}{4}$th measure

F = cumulative frequency up to the interval containing the $\frac{N}{4}$th measure

f = frequency of the interval containing the $\frac{N}{4}$th measure

i = size of the interval containing the $\frac{N}{4}$th measure

As in the formula given for the median, the $\frac{N}{4}$th measure here means the $\frac{N}{4}$th measure counting from the *lowest* values, not counting from the highest as with ranks. As with the median, however, the quartiles can be computed from either direction, if the formula is modified, and the results will be identical. Evaluating the formula for our example from Table 14, we have

$$Q_1 = 2.5 + \frac{\frac{117}{4} - 19}{19} \times 2 = 3.58$$

The formula for Q_3 is

$$Q_3 = l + \frac{\frac{3N}{4} - F}{f} i \qquad (5)$$

where the symbols have the same meanings as in the formula for Q_1 except they refer to the interval containing the $\frac{3N}{4}$th measure. Evaluating the

formula for Q_3 from the data of Table 14, we have

$$Q_3 = 6.5 + \frac{\dfrac{3 \times 117}{4} - 64}{24} \times 2 = 8.48$$

Since Q_2 is identical with the median, already computed by formula (8) on pages 111–112, we shall not recompute it.
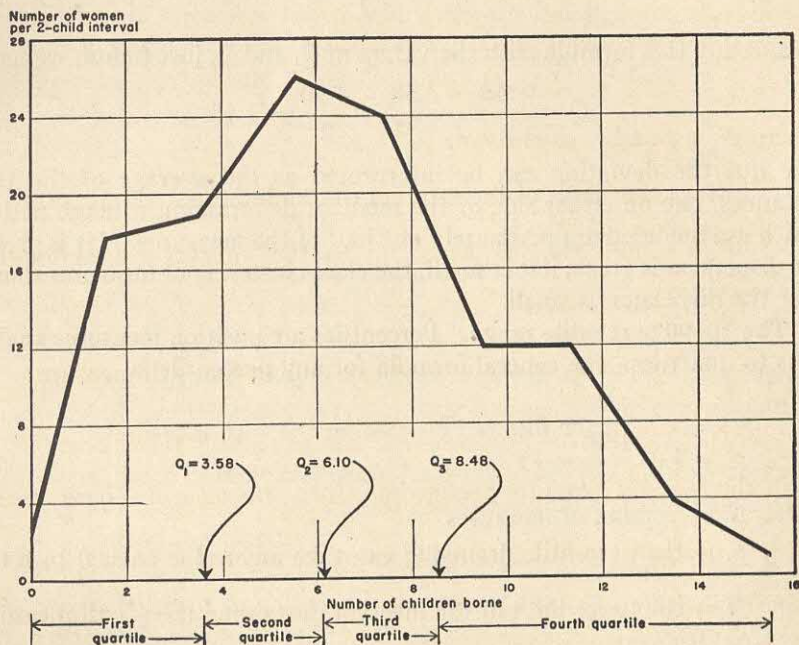
Figure 15. Frequency Distribution with Quartiles Indicated of Number of Children Borne by 117 White Tenant Farm Women, 1938. Coordinate chart with 2-child intervals. (Source: Table 14.)

Figure 15 shows a diagram of the quartile positions and of the parts of the distribution referred to as quartiles. From inspection of the diagram we note that the second and third quartiles cover a smaller range of values than the first and fourth; this is because measures usually cluster around their central value more concentratedly than at their extreme values. The quartiles also tell us that one fourth of this group of women have as many as 8.48 children, that one half of them have between 3.58 and 8.48 children, and that one fourth of them have fewer than 3.58 children.

From these quartiles we construct a summarizing measure of dis-

persion known as the *quartile deviation* or the *semi-interquartile range*. It is defined as the mean distance of Q_1 and Q_3 from Q_2, or by formula,

$$Q = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2}$$

or more simply,

$$Q = \frac{Q_3 - Q_1}{2} \tag{6}$$

Evaluating this formula with the values of Q_1 and Q_3 just found, we have

$$Q = \frac{8.48 - 3.58}{2} = \frac{4.90}{2} = 2.45$$

The quartile deviation can be interpreted as the average of the two distances, one on either side of the median, determining a range within which are included approximately one half of the measures. If it is great, the dispersion is great; if it is small, the close clustering of measures shows that the dispersion is small.

The 10–90 percentile range. Percentiles are position measures analogous to quartiles. The general formula for any percentile is

$$p_i = l + \frac{\frac{jN}{100} - F}{f} \, i \tag{7}$$

where $N =$ number of measures

$p_i =$ the percentile desired (j can take any value from 0 to 100).

$l =$ the lower limit of the interval containing the $\frac{jN}{100}$th measure

$F =$ the cumulative frequency up to the interval containing the $\frac{jN}{100}$th measure

$f =$ the frequency of the interval containing the $\frac{jN}{100}$th measure

$i =$ the width of the interval containing the $\frac{jN}{100}$th measure

As with the computation of medians and quartiles, one must first find the value of $\frac{jN}{100}$ and locate the interval in which the $\frac{jN}{100}$th measure falls. If the 10th and the 90th percentiles are computed, the range between them (within which 80 percent of the measures fall) is sometimes used as a measure of dispersion, chiefly in educational research involving test

scores. Since the 10–90 percentile range is so seldom used in sociological research, we shall not illustrate its computation.

The standard deviation. The standard deviation (from the mean) is the most useful and now the most used of all measures of dispersion. Therefore, the expenditure of time necessary to master the computation and the meaning of the standard deviation will be amply repaid by the resulting facility in understanding the more elaborate summarizing measures based upon it. The *standard deviation*, designated as *s*, may be defined as the square root of the mean of the squared deviations from the mean, or by formula,

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} \tag{8}$$

or since

$$x = X - \bar{X}$$

$$s = \sqrt{\frac{\Sigma x^2}{N}} \tag{9}$$

The standard deviation is a quantity which occurs in the algebraic equations describing many types of distributions and in other equations and formulas for describing the relationship between two or more distributions. Since in more elaborate statistical theory and procedures this measure has to be used, it is well to get thoroughly acquainted with it in simple descriptive statistics.

Computation of the standard deviation with ungrouped data. Whenever we have formulas calling for deviations from the mean, that is, for *x*'s, we try if possible to reduce the formulas to other forms. Preparing the Σx^2 to evaluate in formula (9) involves computing the mean, subtracting the mean from each of the individual measures, squaring each *x* resulting from a subtraction, and adding the squares. The step we especially wish to avoid is the *N* subtractions necessary to get the deviations. By simple algebraic substitutions, with only the process of summation added to the processes learned in elementary algebra, we can change formula (8) so that the subtraction step will be avoided. We begin with formula (8)

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

Expanding the binomial $(X - \bar{X})^2$,

$$s = \sqrt{\frac{\Sigma(X^2 - 2X\bar{X} + \lambda^2)}{N}}$$

Here one must become familiar with some of the simpler rules of summation. One rule is that if we have a summation sign before a polynomial, we can remove the parentheses and sum each term separately, thus,

$$s = \sqrt{\frac{\Sigma X^2 - \Sigma 2X\bar{X} + \Sigma \bar{X}^2}{N}}$$

Only the variable whose value changes, X, is affected by the summation, and therefore constants such as "2" and $\bar{X}$ can be placed in front of the summation sign. In doing this one must remember that in case a term contains no variable, such as the term $\Sigma\bar{X}^2$, the summation sign means simply that we must add as many $\bar{X}^2$ as there are values of the variable to be summed, N in this case. Applying these two rules, we get

$$s = \sqrt{\frac{\Sigma X^2 - 2\bar{X}\Sigma X + N\bar{X}^2}{N}}$$

Breaking the fraction up into three parts we get

$$s = \sqrt{\frac{\Sigma X^2}{N} - \frac{2\bar{X}\Sigma X}{N} + \bar{X}^2}$$

Substituting $\bar{X} = \dfrac{\Sigma X}{N}$, we obtain

$$s = \sqrt{\frac{\Sigma X^2}{N} - 2\left(\frac{\Sigma X}{N}\right)^2 + \left(\frac{\Sigma X}{N}\right)^2}$$

Combining the last two terms, we have

$$s = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} \tag{10}$$

We can write this in different forms, thus,

$$s = \sqrt{\frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}{N}} \tag{11}$$

$$s = \sqrt{\frac{N\Sigma X^2 - (\Sigma X)^2}{N^2}}$$

$$s = \frac{1}{N}\sqrt{N\Sigma X^2 - (\Sigma X)^2} \tag{12}$$

Formulas (10), (11), and (12) are equally acceptable though each has certain advantages. Formula (10) is probably easiest to remember, while formula (12) is simplest to compute with a calculator. Formula (11)

gives Σx^2 as its numerator, and it is frequently desirable to have this term computed separately for further work. These formulas are for ungrouped data only.

From Table 8 we can make the necessary computations for substituting in one of these formulas. We find that $\Sigma X = 742$, $\Sigma X^2 = 6,054$, and $N = 117$.

Substituting in formula (11) we have

$$s = \sqrt{\dfrac{6,054 - \dfrac{(742)^2}{117}}{117}} = \sqrt{\dfrac{6,054 - 4,705.675}{117}}$$

$$= \sqrt{\dfrac{1,348.325}{117}} = \sqrt{11.5241} = 3.395$$

Computation of the standard deviation from grouped data. Although it is possible to compute the standard deviation by using the formula,

$$s = \sqrt{\dfrac{\Sigma f d^2}{N}} \tag{13}$$

where the symbols have the same meanings as before, the short method is so much shorter that it is highly recommended. Table 15 shows the computations necessary for computing the standard deviation by the short method. By comparison with Table 14 it can be seen that for com-

Table 15. COMPUTATION FOR STANDARD DEVIATION FROM GROUPED DATA
OF TABLE 12

Class limits (1)	m (2)	f (3)	d' (4)	fd' (5)	$f(d')^2$ (6)
0	0.0	1	−2.75	− 2.75	7.56
0.5– 2.4	1.5	18	−2	−36	72
2.5– 4.4	3.5	19	−1	−19	19
4.5– 6.4	5.5	26	0	0	0
6.5– 8.4	7.5	24	1	24	24
8.5–10.4	9.5	12	2	24	48
10.5–12.4	11.5	12	3	36	108
12.5–14.4	13.5	4	4	16	64
14.5–16.4	15.5	1	5	5	25
Sums		117		47.25	367.56

$$s = i\sqrt{\dfrac{\Sigma f(d')^2}{N} - \left(\dfrac{\Sigma f d'}{N}\right)^2} = 2\sqrt{\dfrac{367.56}{117} - \left(\dfrac{47.25}{117}\right)^2}$$

$$= 3.45$$

puting the standard deviation by the short method only one column is needed in addition to those used for computing the mean by the short method. The entries in this additional column, labeled $f(d')^2$, are made by multiplying the entry in the fd' column by the corresponding entry in the d' column. The sum of the entries in column (6), the $f(d')^2$ column, divided by N gives the average squared deviation from the *guessed* mean in step deviation units. This average is corrected by subtracting the square of the difference (in step deviation units) between the guessed mean and the true mean, which has already been computed in obtaining the mean by the short method. The square root of the corrected mean is then extracted to obtain the standard deviation in step deviation units. Finally, we multiply by the size of the class interval to obtain the standard deviation in the original units of measurement. All of these processes are described by the formula,

$$s = i\sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \qquad (14)$$

Substituting data from Table 15 in formula (14), we have

$$s = 2\sqrt{\frac{367.56}{117} - \left(\frac{47.25}{117}\right)^2} = 2\sqrt{3.1416 - 0.1631}$$

$$= 2\sqrt{2.9785} = 2 \times 1.7258 = 3.45$$

Variation and variance. Let us consider two concepts closely related to the standard deviation—variation and variance. Although different measures of *variation* have been proposed, the sum of the squares of the deviations of measures from their mean is coming more and more to be used as the measure of total variation of a distribution. In symbols the measure of total variation is Σx^2 or what can be seen is its equivalent from comparison of formulas (9) and (11),

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \qquad (15)$$

The numerical value of this measure of variation in our example is 1,348.325 as can be seen from computations on page 123. This is a very useful concept for later investigation of relationship of two or more characteristics, when we shall learn how to determine what proportion of the variation in one characteristic is associated with the variation in another.

The second concept, *variance*, denoted by s^2, is obtained by dividing the total variation by the number of measures and can be thought of as

a measure of average variation. In symbols the variance is defined as $\frac{\Sigma x^2}{N}$ or its equivalent, thus,

$$s^2 = \frac{\Sigma x^2}{N} = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N} \tag{16}$$

The numerical value of the variance in our example is 11.5241 as can be seen from computations on page 123. Often variance is defined as the square of the standard deviation, but in computation we always arrive at the measure of variance first, and it seems more logical to think of the variance as the average squared deviation and of the standard deviation as the square root of the variance. It also makes the process of computing the standard deviation easier to follow if one thinks of it in three steps: first, we get the sum of the squared deviations, which is the total variation; second, we divide this quantity by N to get the mean variation or variance; third, we take the square root of the variance to get the standard deviation.

Interpretation and use of the standard deviation. Until one has acquired some experience with the standard deviation, it is difficult to get much meaning from it. Of course, it is a measure of absolute dispersion—that is, the more closely the items cluster around a central value, the smaller the standard deviation will be; and the more they spread out over the range, the larger it will be. The most important uses of the standard deviation are to be found in inductive statistics, but we shall suggest here one of the ways in which it is used which does not involve induction. If we wish to compare an individual's measure on two different characteristics, the comparison in terms of the original units of measurement is of little value because the units are different. If measures have been made for a group and the summarizing measures of the distribution of each characteristic worked out, it is possible to reduce any individual's observed measure to standard units which are independent of the original units. In educational research, these are called standard scores or standard measures, designated by the letter z, and defined thus,

$$z = \frac{X - \bar{X}}{s} = \frac{x}{s} \tag{17}$$

For example, let us find the standard measure in the number of children borne for woman with serial number 5, who has 10 children. Substituting in formula (17), we have

$$= \frac{10 - 6.31}{3.45} = +1.07$$

Now the measure of the same woman in number of grades completed (computed in relation to the same group) may be -1.63 (1.63 standard units below the mean). The measures on "fertility" and "educational level" are actually comparable now that they are in standard units, although one was originally measured in terms of number of children borne and the other in terms of number of grades completed. It must be remembered, however, that such standard measures have reference to a group. The summarizing measures, $\bar{X}$ and s, by which an observed measure is transformed into a standard measure, must be obtained from data on the distribution of a characteristic among a *group* of varying units. Such uses of the standard deviation will be found convenient and valuable when one becomes accustomed to thinking in terms of them.

Relative measures of dispersion.   All of the measures of dispersion described so far—the range, the mean deviation, the quartile deviation, the 10–90 percentile range, and the standard deviation—are measures of absolute dispersion, that is, they are given in terms of the original measures. It is obvious that a standard deviation of a certain amount may not indicate so much *relative* dispersion in a group with a high mean as in one with a low mean. For instance, a standard deviation of 3.45 children in our group of high fertility ($\bar{X} = 6.31$) indicates relatively less dispersion than an equal standard deviation in a group of women with a mean of 3.5 children. It is possible to construct a measure of *relative* dispersion by finding what percentage the standard deviation is of the mean. Such a measure is the coefficient of variability, denoted by the symbol V, and computed by the formula,

$$V = \frac{s}{\bar{X}} \tag{18}$$

expressed either as a proportion or a percentage. Substituting the data of our example, we have

$$V = \frac{3.45}{6.31} = .547 \text{ or } 54.7 \text{ percent}$$

whereas in the hypothetical group of lower fertility,

$$V = \frac{3.45}{3.50} = .986 \text{ or } 98.6 \text{ percent}$$

Thus, we see that these two groups (one hypothetical) which have the same measure of absolute dispersion, have quite different measures of relative dispersion. One can construct relative measures of dispersion by computing the percentage any measure of absolute dispersion is of any measure of central tendency, but such measures are not very generally employed.

DESCRIPTION OF THE FORM OF THE DISTRIBUTION

General type of form. The best way of describing simply the form of a distribution is by a graphic presentation of the distribution. For distributions of quantitative characteristics among very great numbers of units where very small class intervals are chosen, either histograms or coordinate charts tend to approach a smooth curve. Therefore, we can think of curves as the limiting case of the form of quantitative distributions when the number of varying units observed becomes infinitely great and the size of class intervals used becomes infinitesimal. We shall call such imaginary curves quantitative distribution curves or frequency curves although they are an abstraction and not an actual graphic representation of any real distribution. The more common forms of quantitative distribution curves can be classified into three types: the *I*-curves,
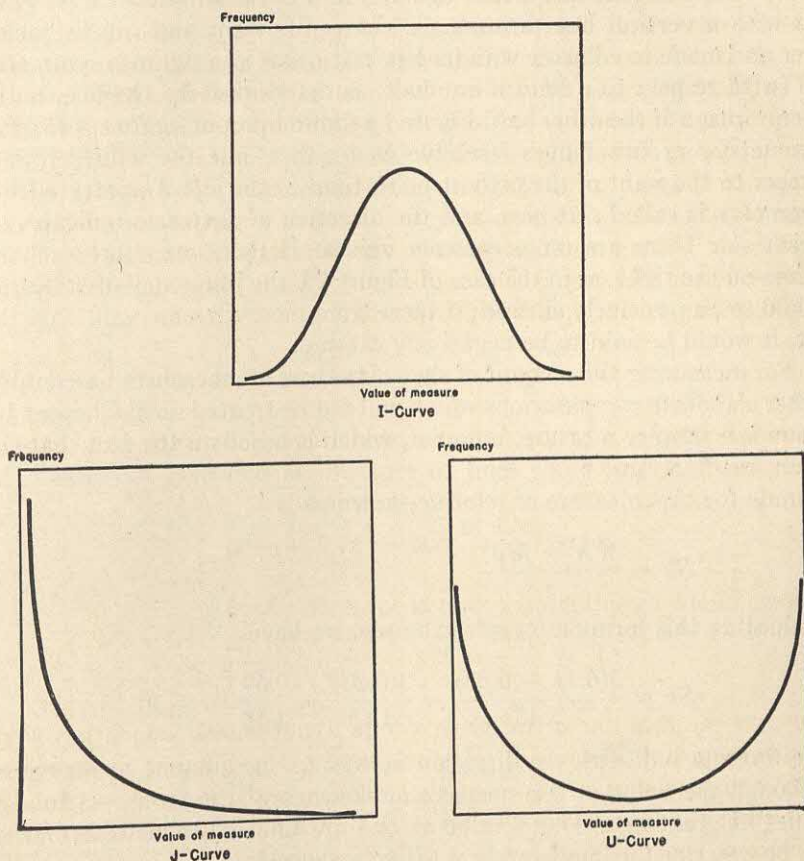
Figure 16.   Types of Frequency Curves.

the *J*-curves, and the *U*-curves. The *I*-curve is one with its greatest frequency (and therefore height) approximately in the center, while its extremities touch or approach the horizontal axis. The *J*-curve is one with its greatest frequency at one end, while its other end touches or approaches the horizontal axis. The *U*-curve has great frequencies at both ends with its smallest frequency in the middle. Figure 16 illustrates these three most common general types. Now by visual comparison of the histogram or coordinate chart of an observed distribution, one can usually see that it resembles one of these curves more than the others and can, therefore, classify the form of the distribution as being of an *I*-type, a *J*-type, or a *U*-type. By comparison of Figures 13 and 16 it can be seen that our example is an *I*-type distribution. By far the greater number of the distributions in which sociologists are interested are of the *I*-type.

Symmetry and skewness. There are other aspects of form to be described besides general type. A figure is said to be symmetrical if, when cut with a vertical line through the center, its right half can be folded over and made to coincide with its left half—that is, a figure is symmetrical (with respect to a line) if one half, as determined by the line, is the mirror image of the other half. Figure 13 shows that our figure approaches symmetry; its two halves resemble each other, but the values extend farther to the right of the highest point than to the left. Departure from symmetry is called *skewness*, and the direction of skewness indicates on which side there are more extreme values. If there are more extreme values on the right, as in the case of Figure 13, the frequency distribution is said to be positively skewed; if there were more extreme values on the left, it would be said to be negatively skewed.

For measuring the amount of skewness there are measures based upon rather elaborate computations which will not be treated until Chapter 14. There is a simpler measure, however, which is based on the fact that the mean, median, and mode tend to separate as skewness increases. The formula for this measure of relative skewness is

$$Sk = \frac{3(\bar{X} - Md)}{s} \tag{19}$$

Evaluating this formula for our example, we have

$$Sk = \frac{3(6.31 - 6.08)}{3.45} = \frac{3(0.23)}{3.45} = \frac{0.69}{3.45} = +.20$$

The formula indicates the direction as well as the amount of skewness. Although the value of this measure of skewness varies from -3 to $+3$, values as great in absolute value as one are unusual.[2] Our distribution can be described as moderately positively skewed.

---

[2] Frederick E. Croxton and Dudley J. Cowden, *Applied General Statistics* (New York: Prentice-Hall, 1939), pp. 251–253.

There is another aspect of form, *kurtosis*, which refers to the peakedness or flatness of the curve representing a frequency distribution. This measure is also based upon more elaborate computations, and treatment of them will be delayed until the chapter on the normal curve (Chapter 14).

EXAMPLE OF THE DESCRIPTION OF A DISTRIBUTION

To make clear the procedures most frequently used in describing quantitative distributions, we shall illustrate the computations of the most important summarizing measures by an additional example.

Table 16. FREQUENCY DISTRIBUTION OF THE 100 COUNTIES OF NORTH CAROLINA BY CLASS INTERVALS OF PERCENTAGE CHANGE IN POPULATION, 1940–1950, AND COMPUTATION FOR MEAN, MEDIAN, MODE, AND STANDARD DEVIATION

Percentage change in population 1940-1950 (1)	Fre-quency f (2)	Midpoint of class interval m (3)	Step deviation d' (4)	fd' (5)	$f(d')^2$ (6)	F (7)
$-25.0 - -15.1$	2	-20.05	-3	-6	18	2
$-15.0 - - 5.1$	8	-10.05	-2	-16	32	10
$-5.0 - 4.9$	32	-0.05	-1	-32	32	42
$5.0 - 14.9$	38	9.95	0	0	0	80
$15.0 - 24.9$	9	19.95	1	9	9	89
$25.0 - 34.9$	7	29.95	2	14	28	96
$35.0 - 44.9$	1	39.95	3	3	9	97
$45.0 - 54.9$	1	49.95	4	4	16	98
$55.0 - 64.9$	1	59.95	5	5	25	99
65.0 and over	1	133.3 [a]	12.34	12.34	152.276	100
Sums	100			-6.66	321.276	

$$\bar{X} = \bar{X}' + \frac{\Sigma fd'}{N} i = 9.95 + \frac{-6.66}{100} 10 = 9.28$$

$$Md = l + \frac{\frac{N}{2} - F}{f} i = 4.95 + \frac{50-42}{38} 10 = 7.055$$

$$M_o = l + \frac{\Delta s}{\Delta s + \Delta g} i = 4.95 + \frac{6}{6 + 29} 10 = 6.66$$

$$s = i\sqrt{\frac{\Sigma f(d')^2 - \frac{(\Sigma fd')^2}{N}}{N}} = 10\sqrt{\frac{321.276 - \frac{(6.6)^2}{100}}{100}} = 17.91$$

[a] Since this is an open-end interval, we use the actual value of the case.
Source: *1950 Census of Population*, Preliminary Counts, Series PC-3, No. 4, "Population of Counties: April 1, 1950."

Table 16 shows the distribution of the 100 counties of North Carolina according to the percentage change in population, 1940 to 1950. In this table the last class interval is an open-end interval, *65.0 and over*. Since the midpoint of an open-end interval cannot be determined, we have recorded as the midpoint of this interval the actual percentage increase of the one county falling in this interval. If more than one county fell into this interval, we would have recorded the mean of the values in the interval. We have not recorded the actual percentage increase in each of the 100 counties, but these figures may be found in the original source of Table 16.

Computations of the mean and standard deviation. It is possible to compute these two measures either from grouped or ungrouped data. In order to compute these measures from ungrouped data, it is necessary to have ΣX and ΣX^2. In computing ΣX the fact that some of the values are positive and some are negative must be taken into consideration. The values of the two quantities are as follows:

$$\Sigma X = 928.3$$
$$\Sigma X^2 = 38,792.93$$

Computation of the mean and standard deviation is as follows:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{928.3}{100} = 9.283$$

$$s = \frac{1}{N}\sqrt{N\Sigma X^2 - (\Sigma X)^2} = \frac{1}{100}\sqrt{100\,(38,792.93) - (928.3)^2}$$
$$= 17.371$$

The computation of these same two measures from grouped data is shown in Table 16. In general, the values of summarizing measures computed from grouped and ungrouped data will be slightly different.

In this situation where we have one measure that is so widely different from all the other measures, being more than twice as great as the next largest measure, it is interesting to see the effect of one such extreme case. If this case is omitted from the ungrouped data we have:

$$N = 99, \quad \Sigma X = 795.0, \quad \Sigma X^2 = 21,024.04$$
$$\bar{X} = \frac{795.0}{99} = 8.03$$

$$s = \frac{1}{99}\sqrt{99\,(21,024.04) - (795.0)^2} = 12.16$$

The inclusion of this one extreme case increases the mean by nearly 16 percent and increases the standard deviation by 43 percent. The student

should verify for himself the fact that the mode is unaffected by the in-
clusion or exclusion of this extreme case and the median is changed
negligibly.

SUMMARY AND COMPARISONS OF SUMMARIZING MEASURES AND OTHER DEVICES FOR DESCRIBING QUANTITATIVE DISTRIBUTIONS

We have defined and illustrated summarizing measures and other de-
vices for describing a quantitative distribution as a whole and three of its
specific aspects—its central tendency, its dispersion, and its form. We
have suggested several different measures for describing most of the
features of a quantitative distribution and usually several different ways
of computing each measure. Now the analysis and description of any one
distribution does not necessitate the computation of all the measures sug-
gested by every method of computation explained. To help the reader
choose the measures and methods most appropriate for his problem, we
shall give a summary of the measures and other devices and of their
methods of computation, comparing alternate measures and methods and
pointing out the advantages and disadvantages of each.

Methods for describing the distribution as a whole. The four methods
for describing a frequency distribution as a whole are: (1) an *unordered
listing* of the measures, (2) an *array*, (3) a *frequency table*, and (4) a
graphic form. For a single distribution it is hardly probable that all four
of these methods would be presented in a report although any one or any
combination of the four may be used. Let us consider them in order.

The *unordered listing* of measures is almost never used to present
findings in published form except in the case where the number of indi-
viduals whose measures are listed is small (approximately 100 or fewer)
and where there is an established conventional order of listing the indi-
viduals. For instance, in a report where the 48 states are studied in re-
gard to various characteristics, it may be preferable to list the states
always in alphabetical or some geographical order, regardless of the
magnitudes of their measures.

The use of the *array* is likewise restricted to cases where the number of
individuals is not great (again approximately 100 or fewer). The array
is most commonly employed where the number of individuals is small
enough and the nature of the individuals is such that interest attaches to
the measure and ranking of individual units. If the identity of the indi-
vidual unit which has a certain measure on a characteristic is to be empha-
sized, an array including names, serial numbers, or some other identifying
device is appropriate. A modification of the array is a list of the individuals
in order of their measures with the actual values of the measures omitted

and the ranks of the measures used instead. This method is also often used with the 48 states when data are presented on only one characteristic at a time. In case the number of individuals is quite small, 30 or fewer, and a frequency table or graphic form is not appropriate because of the small number of cases, an array is the preferred method of describing the distribution as a whole.

The *frequency table* is the best method of presenting fairly detailed information on the distribution of a measurable characteristic among a large number of individual units. It is to be chosen in preference to the array if the number of cases is so great that the array would be unwieldy, or if the number is simply great enough to prevent the frequencies of classes from showing marked irregularities, so long as the identification of the individuals is not essential. Whether or not the frequency table is to be used as a presentation form, a frequency table must be prepared if a graphic form of the distribution is to be made or if the computations are to be made by the methods for grouped data.

A *graphic form* of a frequency distribution should be used to describe the distribution as a whole, when detailed information is not necessary or when the form of the distribution needs to be described also. Its usefulness is limited to cases where there are enough individuals to give a fairly regular distribution by intervals.

Choice of using grouped or ungrouped data. Before beginning the analysis of a quantitative distribution, it is necessary to decide whether to compute the summarizing measures by the methods of grouped or ungrouped data. The first consideration is that of number of cases and the decision is easy if the number is very great or very small. If the number is 500 or more, grouping is indicated; if the number is 30 or fewer, grouping is not advised. But if the number is somewhere between these two limits, it is more difficult to decide whether grouping should be done or not. If there are no marked irregularities in the data, a rather safe rule is that if criteria (1) and (2) listed on page 89 of Chapter 8 can be fulfilled, the number of cases is great enough to justify grouping. Fortunately, this is about the equivalent of the rule that if there are enough cases to make grouping save time in computation, grouping may be used. It will, of course, take practical experience to learn to judge just where the line of demarcation is. A second consideration is the degree of accuracy required in the analysis. Grouping always sacrifices some accuracy and if there is doubt on this score, it is safer to use ungrouped data. A third consideration is that of the further analysis to be made and whether it can best be based on measures computed from grouped or ungrouped data. On this score also it is safer to use ungrouped data in borderline cases, since otherwise further analysis will be restricted to using the identical grouping already chosen, which may not be the best grouping

for more elaborate treatment. Finally, to offset somewhat these last two considerations, if a graphic presentation form is to be used, the data will have to be grouped anyway, and even more time will be saved by using the grouped data. Therefore, unless there is some definite reason for not using computations based upon grouped data, it is usually expedient to use them when a histogram or coordinate chart is to be made of the distribution.

Measures of central tendency. The *arithmetic mean* is the most generally preferred measure of central tendency and should always be computed in any thorough analysis and description of a distribution unless there is some special reason which makes the mean impossible to compute or misleading. The mean has the following properties which are usually considered as advantages in its use as a measure for describing central tendency: (1) it is based upon the values of *all* the observations; (2) it is algebraically defined and therefore can be used as a basis for further analysis; (3) it is more "stable" than other measures (a property which is important especially if the description of the distribution is to be the basis for induction; the meaning of "stable" will be clarified in Part III). The first property listed, however, may be disadvantageous in two cases—when one does not know all the values of the observations and hence cannot compute the mean; or when a few values of the measures may be so extreme that they affect the mean sufficiently to make it unrepresentative of the other values. An example of the first case is where one seeks to analyze a distribution for which his data are given in grouped form with one or two open-end intervals, indicated by such class limits as "$3,000 and over," or "fewer than 10." In such distributions we do not know what the midvalues of the open intervals are and hence cannot compute the mean because it is based upon *every* value in the distribution. Of course, in such a case it is sometimes possible to use an estimate of the midvalue of the interval. For instance, if the last class in an age distribution is given as "80 and over," one may be able to find a life table approximately appropriate to the group whose ages are listed and from it read off the average years of life remaining to those who survive to age 80. This number of years would probably be around six which can be added to the lower limit to determine an approximate midvalue of 86 for the interval. The mean could then be computed approximately. If midvalues of open intervals cannot be satisfactorily estimated, however, other measures of central tendency must be used.

The most commonly encountered illustration of the second case where the mean should not be used is in the distribution of wealth or income or other characteristics associated with wealth or income, for these characteristics have very skewed distributions. If one or two individuals in a group have such extremely high incomes that they raise the mean for the

group above the level of income of all the other individuals in the group, the mean is no longer a good measure of central tendency of the group, for it certainly is not representative of the majority of the group. In other cases of very skewed distributions, similar but less extreme than this, the mean may not be the best measure.

The *median* is characterized by the following properties: (1) it is based upon the values of only the central measures, although it is affected by the positions relative to the median (above or below) of the extreme ones; (2) it is not algebraically manipulable to the same degree that the mean is; (3) it is usually between the mean and the mode in stability. The median is appropriate as a measure of central tendency whenever the effects of the extreme measures are to be minimized as in the income illustration above or in any very skewed distribution. If in a series of values there are one or two which vary so widely from the rest that one is suspicious of the accuracy of their recording, but cannot check it, the median may well be used, since it will not be so badly affected as the mean if the suspicious values are wrong. Often the median is used along with the mean to show symmetry or skewness by a comparison of their values.

The *mode* is characterized by the following properties: (1) it is based upon only the values of the central measures, neither values nor positions of extreme measures affecting it at all; (2) it is not algebraically defined or manipulable (unless one actually fits a curve to the distribution by elaborate computation); (3) it is less "stable" than either the mean or the median. The mode is the appropriate measure of central tendency to use when we wish to ignore altogether the effect of extreme values, when we are concerned with only the most common or typical value. Because of its instability the mode should not be used unless there are enough cases to afford a fairly regular distribution around the interval of greatest frequency.

When a distribution is symmetrical, the mean, median, and mode are identical. When a distribution is skewed, they are not identical—the mean is farthest in the direction of the skew, the median next, and the mode next. Therefore, either the median or the mode or both may be computed to be used along with the mean to demonstrate skewness. If there is a fairly large number of cases and if the distribution is only slightly or moderately skewed, the following relation holds approximately,

$$\bar{X} - Mo = 3(\bar{X} - Md) \tag{20}$$

The relation is shown schematically in Figure 17. The relation can be used to check roughly on the accuracy of the determination of the several measures of central tendency. Let us check the values obtained from the ungrouped data on number of children of the 117 white tenant farm women. The measures of central tendency are as follows:

$$\bar{X} = 6.34, \; Md = 6.17, \; Mo = 6.00$$

Now,

$$\bar{X} - Mo = 0.34$$
$$3(\bar{X} - Md) = 3(0.17) = 0.51$$

Thus we see that in our example the difference between the mean and the mode is only twice instead of three times as great as the difference be-

Frequency
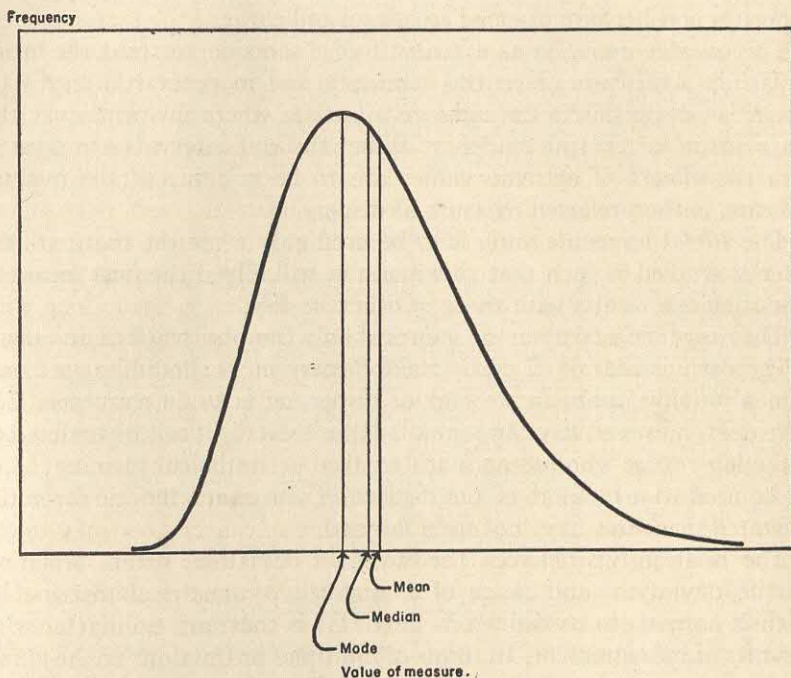
Figure 17.   Relative Positions of the Mean, Median, and Mode in a Moderately Skewed Distribution.

tween the mean and the median. Yet even with so small a number of cases as 117 and with a distribution as skewed as that of our example, the relation is approximated by these three measures of central tendency and at least their relative positions are checked.

Measures of dispersion. The *standard deviation* among measures of dispersion is comparable to the mean among measures of central tendency —that is, it is the most important and should be used to describe dispersion unless there is a reason for not doing so. Corresponding to the mean, it has the following properties: (1) it is based upon all values of the observations; (2) it is algebraically defined and adapted for further analysis; (3) it is more stable than the other measures of dispersion. Since, like the mean, it is based upon all values of the observations, the standard devi-

ation cannot be computed unless they are all known, and, like the mean, it has to be replaced by some other measure when they are not known. The standard deviation has meaning even if a distribution is badly skewed, but often another measure of dispersion is preferable in such a case.

The *mean deviation* shares the first property with the standard deviation, but it is not advised because it does not share the second and third properties and because of other reasons listed earlier.

The *quartile deviation* as a measure of dispersion corresponds to the median as a measure of central tendency and in general is used as a measure of dispersion in the same distributions where the median is used as a measure of central tendency. When the end intervals are open or when the effects of extreme values are to be minimized, the quartile deviation is the preferred measure of dispersion.

The *10–90 percentile range* is to be used only when the nature of the material studied is such that this measure will afford the best means of comparing the results with those of other studies.

The *range* is based upon the values of only two observations and hence is the most unstable of all measures of dispersion. It should be used only when a roughly approximate idea of dispersion is to be conveyed. The range does, however, have two great advantages: (1) it can be understood by the lay reader who has no mathematical or statistical training; (2) it can be used as a measure of the dispersion of a characteristic for which calibrated measures have not been devised.

The relationship between the standard deviation, mean deviation, quartile deviation, and range of a perfectly symmetrical distribution which is normal (to be defined in Part III) is constant, no matter what the units of measurement. In terms of multiples or fractions of the standard deviation, the important measures of dispersion are as follows:

$$Q = 0.6745s$$
$$MD = 0.7979s$$

The relationship between the range and the standard deviation is a function of the number of cases as shown below:

N	$\dfrac{R}{s}$
30	4.09
100	5.02
500	6.07

These relations make possible a rough estimation of one measure of dispersion when another is known. They can also be used for roughly checking results. For instance, the range in the number of children of the 117

women was 16. An estimate of the standard deviation from the knowledge that it is about one fifth of the range is obtained thus,

$$\frac{16}{5} = 3.2$$

which is fairly close to the computed value of 3.39. This sort of quick check is especially useful for locating gross mistakes, or "decimal point" mistakes. Since in a normal distribution the distance from one quartile deviation below the mean to one quartile deviation above the mean includes one half of the measures of the distribution, a rough estimate of the range including one half the measures can be made in a problem where only the mean and the standard deviation have been computed by evaluating the expression,

$$\bar{X} \pm 0.6745s$$

Methods of describing form. Until the more elaborate measures of skewness and kurtosis are taken up, there is not much choice of devices for describing the form of a frequency distribution. An experienced statistician can get a good idea of the form of a distribution from close study of a frequency table, but a graphic presentation of the distribution makes the form evident to all. If elaborate analysis is not to be performed, a histogram or line chart of the frequency distribution usually suffices for description of form of the distribution, although the coefficient of skewness suggested may be used to supplement the chart.

SUGGESTED READINGS

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chaps. VIII, IX, and X.

Dixon, W. J., and Massey, Jr., F. J., *Introduction to Statistical Analysis* (New York: McGraw-Hill, 1951), Chaps. 2, 3, 5, 6.

McNemar, Quinn, *Psychological Statistics* (New York: Wiley, 1949), Chap. 3.

Peatman, J. G., *Descriptive and Sampling Statistics* (New York: Harper, 1947), Chaps. 5, 6, and 7.
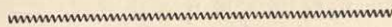
Scales and Indexes

Content of chapter. In sociological research the terms "scale" and "index" are used to refer to all sorts of measures, absolute or relative, single or composite, usually indirect or partial, sometimes patterned after the economist's index number and sometimes after the psychologist's and educationist's tests and scales. Since so many of the characteristics in which sociologists are interested are not simply, directly, and absolutely measurable, the increasing use of such scales and indexes promises to be fruitful. We shall first take up some characteristics of indexes and then one type of scale construction, namely, that developed by Louis Guttman. After illustrating the construction of a Guttman scale, we will discuss types of areas that are not likely to be scalable and will discuss and illustrate some of the simpler methods of construction of arbitrary indexes developed by economists, sociologists and others. More elaborate methods of index construction will be taken up in Chapter 26.

THE NATURE AND TYPES OF INDEXES

Definition and illustration of terms. For the purpose of the following discussion we shall define "index" as follows: "An index is one or a set of measures for one or a group of units which is used to measure *indirectly* the incidence of a characteristic that is not directly measurable." Thus, the percentages of illiterates in a series of demographic units may be used as negative indexes of the general cultural level of the units which cannot be measured directly. The number of rooms in the dwelling of a family may be used as an index of the family's economic status. The number of correct answers on a certain test may be used as an index of the subject's intelligence. The feature by which we are distinguishing an index from other measures is its *indirectness* in giving information about the characteristic being studied, in contradistinction to the *directness* of measures such as age and number of children borne. Of course, the direct-indirect division of measures is not an absolute dichotomy; there are borderline

measures which would be difficult to classify. But the distinction seems useful in the clarification of methods. It must be noted also that what we shall here classify as indirect measures or indexes of one characteristic are direct measures of one or more other characteristics which are either parts of the first more complex characteristic or are related to it in some way.

Now let us distinguish between *simple* and *composite* indexes. A *simple* index is one or a set of measures of a single measurable characteristic, which is used as an indirect measure of another not directly measurable characteristic; a *composite* index is one or a set of measures, each of which is formed by combining simple indexes. If simple indexes of the level of living, such as number of income tax returns, number of radios, and number of telephones per person are combined by some one of the ways explained below into one figure, this figure is a composite index of the level of living. An intelligence test score is actually a composite index of intelligence since numerical values for performance on each item of the test are combined to give the test score. In construction of an index of general ability, however, it may be more convenient to refer to an intelligence test score as one of the simple indexes to be combined into a composite index.

An index, either simple or composite, may be expressed in the original units of measurement or as a ratio of two measures, such as "number of radios per 1,000 population." (The first is more common where the "measure" is of a relatively elemental unit, the second where the measure is of a composite unit such as a county or an institution.) It is usually the aim of index makers, however, to transform the original measures into some standard form which will facilitate the process of comparison of the measures of different units. The educationist may arrange his scoring so as to express the test scores of individuals in terms of a group norm. The economist usually, although not always, relates his index to what he considers a "normal" year and expresses the index for other years as percentages of this. The sociologist, like others who make indexes for series of units which vary in some aspect other than time, may express his index as a percentage of the average for the whole group or of some one unit he has reasons for selecting as a base.

Now when a composite index relating to economic data has been expressed in terms of the index value of the base year or unit, it is called an "index number." When a composite index relating to human abilities has been administered to some standard group and checked on several criteria, the methods by which the data for the index are gathered, along with the methods for combining information and determining a score, are called a "standardized test." If a similarly developed index is designed for the indirect measurement of human characteristics other than

"abilities," as in the case of indirect measurement of attitudes, it may be called a "standardized scale." Sociologists have often called their indexes "index numbers" if their methods of construction were similar to those of the economists; or they have called them "tests" or "scales" if they were similar to the methods of the psychologists and educationists; and even more often they have simply used the generic term, "index." The Guttman scale, which we shall take up, is a special type of composite index.

Basic criteria: validity and reliability. Whether a simple index is to be used alone or in combination with other indexes to form a composite index, the two basic criteria for its selection are the same as those for any measure or measuring instrument—validity and reliability. These criteria might have been considered earlier, but they are so much more of a problem in indirect than in direct measures that it seems best to treat them in their particular relation to indexes since their application to direct measures is simpler. In index construction these two terms, validity and reliability, have rather specialized meanings defined by processes involving correlation. In nontechnical language, however, a measuring instrument is said to have *validity* if it measures what it purports to measure; it is said to have *reliability* if it gives the same results consistently. These criteria have to be considered both at the stage of selection of simple indexes and at the stage of evaluating the final composite index after the simple indexes have been combined.

Validity. The validity of a measuring device is usually studied by comparing the results or measures obtained from it with those obtained by another device, the validity of which is already established for measuring the same characteristic. If such a measuring device of established validity is not available—and this is often the case—the problem of establishing validity becomes difficult. Judgment of experts and internal consistency are used to establish the validity of the simple indexes and correspondence with the nearest approximations to measures of the characteristic is used to establish the validity of the composite index.

In case of direct measures, validity is self-evident. In fact, we call those measures direct which unquestionably measure precisely what we intend them to. In the case of indirect measures or indexes, validity is only approximate. This limitation of indexes must be kept constantly in mind, especially when making interpretations of findings. The two sorts of approximations to validity found in indexes are the use of indexes which measure a part of a larger complex being studied, and the use of those which measure something which is not actually a part of the complex studied but is associated with the complex and likely to be present in varying degrees corresponding to the degrees of incidence of the complex. For example, the index of wholesale prices made by the United States Bureau of Labor Statistics combines data on 2,000 commodities or

series.[1] The price of each one of these commodities has what we are calling *partial validity* in indicating the general level of wholesale prices, since it is actually a measure of one part of the complex called "wholesale prices." On the other hand in most composite indexes or scales of socio-economic status, the presence of certain physical equipment, such as a sofa in the living room, is used as an index or reflector of status. A sofa is not a part of status, and the possession of a sofa is not a direct measure of status, but the possession of a sofa tends to be associated with higher degrees of status than nonpossession of a sofa. It is what we call a measurable correlative of a complex which is not directly measurable, and it has what we are calling *correlative validity* as an index of status.

On the whole, economists, who deal largely with phenomena which can be reduced to the common denominator of dollars and cents, do not encounter great difficulty in establishing the validity of their simple indexes; although after these have been combined by rather elaborate methods the validity of the composite index must be examined. But in the fields of psychology and sociology, where the concepts of what one is trying to measure are often vague, the problem of validation of an index is exceedingly difficult at times.

Reliability. A variety of fairly elaborate techniques have been developed, especially in psychology and education, for studying the reliability of a measuring device. If a test is a composite one, the closeness of correspondence between the scores on a test and a retest with an alternate form of it, or between the scores made on different parts of it, is used to indicate the degree of reliability of the test. In all fields careful specifications of definitions and explicit instructions for gathering data in as uniform manner as possible are essential for insuring the reliability of the simple indexes. This is easier for the economist than for those dealing with human responses. If the simple indexes for an economic index number are matters of record, such as tons of pig iron produced in a certain period, all the economist has to do is to check his sources for their integrity and accuracy. If the simple indexes for a socio-economic index or scale are such measures as the presence or absence of a sofa, the index maker will need to be careful of his definitions and instructions to enumerators in order that a day bed will not be listed as a sofa on one occasion and not so listed on another. In a test construction where human responses of the subject are involved, not only must the situation be uniform for the administration of the test, but also the attainment of reliability must be checked by experimental trials using such methods as the "test-retest" or "split-half" methods referred to above.

---

[1] Edgar I. Eaton, "A Description of the Revised Wholesale Price Index," *Monthly Labor Review*, February 1952.

Combination of simple indexes into a composite index. This phase of index making overlaps the phase of selection of simple indexes since that selection is often partly determined by how they are to be combined. In economic indexes where simple indexes are usually partial measures, they may be considered as a *sample* of the possible part measures and, therefore, should be both adequate (sufficient in number) and representative, two criteria of samples which will be discussed in Part III. Often weighting factors are used in the process of combination to make the choice of simple indexes proportionately representative. For instance, in a cost-of-living index, the food prices used might be weighted by the average percentage food expenditures are of total expenditures, and clothing, transportation, and other prices similarly. The economist has many ways of combining his simple indexes. Some of the methods express the simple indexes in relation to the base before combining them; some combine the indexes and then express them in relation to the base. Arithmetic, geometric, and harmonic means, weighted and "unweighted," are used in the various combining methods.

The psychologist or educationist also has a variety of ways in which he combines measures on "items" (simple indexes) into test scores (composite indexes). He may perform his weighting by the number of each type of item he selects, using more items of a type he considers important, or he may in his scoring instructions assign more weight to single items or groups of items. After weights have been assigned, however, the process of combination is usually additive although there are numerous methods employed for arriving at final scores.

FIELDS OF APPLICATION OF METHODS OF INDEX CONSTRUCTION IN SOCIOLOGY

Indexes are widely used in sociological work as is evidenced by the following sampling of titles of various journal articles:

Gough, Harrison G., "A New Dimension of Status: I. Development of a Personality Scale," *American Sociological Review*, 13 (August 1938), pp. 401–409.

Hagood, Margaret Jarman, "Construction of County Indexes for Measuring Change in Level of Living in Farm Operator Families, 1940–45," *Rural Sociology*, 12 (June 1947), pp. 139–150.

—— and Eleanor Bernert, "Component Indexes as a Basis for Stratification in Sampling," *Journal of the American Statistical Association*, 40 (September 1945), pp. 330–341.

Jahn, Julius A., Schmid, Calvin F., and Schrag, Clarence, "The Measurement of Ecological Segregation," *American Sociological Review*, 12 (June 1947), pp. 293–303.

Jahn, Julius A., "The Measurement of Ecological Segregation: Derivation of

an Index Based on the Criterion of Reproducibility," *American Sociological Review*, 15 (February 1950), pp. 100–104.

McMillan, Robert T., "Comparison of Farm Housing Indexes for Oklahoma," *Social Forces*, 24 (December 1945), pp. 174–180.

Porterfield, Austin L., "Rank of the States in Professional Leadership and Social Well-Being," *Social Forces*, 25 (March 1947), pp. 303–309.

Schuessler, Karl, and Strauss, Anselm, "A Study of Concept Learning by Scale Analysis," *American Sociological Review*, 15 (December 1950), pp. 752–762.

Shapiro, Gilbert, "Myrdal's Definitions of the 'South': A Methodological Note," *American Sociological Review*, 13 (October 1948), pp. 619–621.

Sletto, Raymond F., "Index Numbers for Social Security Program Analysis," *American Sociological Review*, 12 (August 1947), pp. 424–429.

Svalastoga, Kaare, "An Index of International Security," *American Sociological Review*, 15 (October 1950), pp. 668–672.

Williams, Josephine J., "Another Commentary on So-Called Segregation Indices," *American Sociological Review*, 13 (June 1948), pp. 298–303.

Because of extensive treatment in other volumes no articles covering the measurement of attitudes are listed above. However, the methods discussed in this chapter, especially the Guttman scale, are applicable to this area of research. The measurement of level of living or socio-economic status is discussed in Chapter 26 since the methods of that chapter are the ones most recently applied to the measurement of level of living.

Since so many variables that are of interest to sociologists are not directly measurable, it is expected that the use of indexes will become of increasing importance in many areas of sociological research.

THE GUTTMAN SCALE

The method of index construction which we shall first take up is known as the Guttman technique of scale analysis.[2] The definition of a scale used by Guttman requires that when the individuals are scored on the scale, their responses to the items making up the scale should be reproducible from their scores. This is the definition for a Guttman scale, and we can illustrate this principle with a simple scale measuring mathematical training. Let the items on our scale be as follows:

1. $14 + 37 =$
2. $53 - 47 =$
3. $\dfrac{3}{4} - \dfrac{5}{7} =$

---

[2] The theory of scale construction given in this chapter is taken from chapters by Guttman and Suchman in *Measurement and Prediction*, by Samuel A. Stouffer, Louis Guttman, Edward Suchman, Paul F. Lazarsfeld, Shirley A. Star, and John A. Clausen (Studies in Social Psychology in World War II, Vol. IV; Princeton, Princeton University Press, 1950). This volume is one of the best and most complete treatments of scaling available at this time, and the reader is referred to it for additional material.

$$4. \; 3X + \frac{7}{8} = \frac{29}{24}; \; X =$$

$$5. \; 3X^2 + 6X = 7; \; X =$$

Persons taking this test will get either 0, 1, 2, 3, 4, or 5 items correct and can, therefore, be assigned one of six scores. Those getting five right scoring one, four right scoring two, etc. down to score six with none right. A person who made a score of four, having gotten two items correct, would in general have gotten the first two items correct and missed the last three since the items are of increasing difficulty. Since being able to work any problem implies the ability to work all preceding problems, this fits our requirement for a scale—that the response made by a person to each item is reproducible from a knowledge of that person's score. Thus, a person who makes a score of two will have gotten the first four items correct and missed the fifth item.

In a similar manner certain attitude areas are scalable. Scalability in this sense implies that the scale is measuring just one variable, or mathematically speaking, just one dimension. Frequently a scale will be measuring more than one variable, and the results are likely to be confusing. For example, it was found that morale could not be measured with a single scale because it is a combination of several variables (or dimensions).[3] However, areas, such as morale, that are not directly scalable can frequently be broken down into scalable subareas.

ILLUSTRATION OF THE GUTTMAN TECHNIQUE

Area of investigation and items to be used. We shall illustrate the Guttman technique of scale analysis with a set of items which relate to duties and functions assigned by law to state boards of education. These items form a scale which measures the legal authority of state boards of education. The need for such a scale might arise if one were investigating the relationship between legal authority of state boards of education and some other variable, such as proportion of tax money spent for education. The need for such a scale might also arise if one is studying legislative and administrative practices of state governments. An area such as this is used for illustrative purposes since there are already available many examples of the application of the Guttman technique in attitude and opinion work.[4]

The illustration used here differs from most applications of the Guttman technique in that the units are areas rather than individuals. All of

---

[3] Samuel A. Stouffer and others, *The American Soldier: Adjustment During Army Life*, Studies in Social Psychology in World War II, Vol. I (Princeton: Princeton University Press, 1949), Chaps. 3, 4, 5.

[4] See Samuel A. Stouffer and others, *Measurement and Prediction, op. cit.*

the ramifications of the application of the Guttman technique to units of this sort have not been investigated, but the results seem to be fruitful.

The data used in constructing the scale of legal authority of state boards of education were collected by The Council of State Governments in a study of state school systems [5] and were not collected with the idea of scaling them. The original data are shown in Table 17.

Construction of scale. The simplest way to test a set of items, such as those of Table 17, for scalability in the Guttman sense is with the use of a scalogram board.[6] However, scalogram boards are not generally available. The technique which will be described here is known as the Cornell technique.[7] In this method the first step is to select the "most positive" reply to each item. The questions in Table 17 are dichotomous since a school board either does or does not have authority in a particular area. We can select either response as the "positive" response but must consistently use the same type of response as positive for all questions. In this case we have taken the existence of legal authority in an area as the positive response. A positive response is assigned a score of one, and a negative response is assigned a score of zero. If the items are not dichotomous, the responses are assigned integral scores beginning with zero for the most negative response. Thus an item having three possible responses would have a score of two for its most positive response.

Having assigned a score for each possible response to each item, we next compute the total score for each individual. This is done for an individual by adding up the scores corresponding to his responses. The individuals are then arranged in order of their total scores from highest to lowest.

In Table 18 the states are shown arranged in order of decreasing score with the score for each state shown in parenthesis. There is a column of Table 18 for each possible response to each question. The first of each pair of columns shows the positive responses, and the second of each pair of columns shows the negative responses. If any item had more than two possible responses, there would be a column for each type of response with the columns for each item arranged in order from positive to negative.[8] The items are arranged in order from the one with the fewest positive responses to the one with the most positive responses. The number of positive responses to each item is shown in the last line of Table 18.

---

[5] *The Forty-Eight State School Systems.* (Chicago: Council of State Governments, 1949) Appendix Table 10, p. 184.

[6] Samuel A. Stouffer and others, *Measurement and Prediction, op. cit.* Chap. 4.

[7] Louis Guttman, "The Cornell Technique for Scale and Intensity Analysis," *Educational and Psychological Measurement,* 7 (1947) pp. 247–279.

[8] It is not essential that we be able to order the responses from positive to negative. See Stouffer *et al., Measurement and Prediction, op. cit.,* p. 102. In general, however, it is possible to order the responses on an a priori basis.

Table 17. EXTENT TO WHICH CERTAIN DUTIES AND FUNCTIONS ARE AS-
SIGNED BY LAW TO STATE BOARDS OF EDUCATION, 1947–48 [a]

State	Duties and functions as described in right-hand column												Duties and functions of state boards of education as indicated in columns 1–12
	1	2	3	4	5	6	7	8	9	10	11	12	
Ala..............	x	x	x	x	x	x	x	x	x				1. Determination of educational poli-
Ariz..............		x	x		x			x	x				cies (32) [b]
Ark..............	x	x	x	x	x	x	x	x	x				2. Adoption of rules and regulations
Calif..............	x	x	x	x	x	x	x	x	x	x		x	which have the effect of law (37)
Colo..............				x	x			x		x			3. Prescription of minimum standards
Conn..............	x	x	x		x	x	x	x	x				in specified areas (34)
Del..............	x	x	x		x	x	x	x	x				4. Determination of regulations gov-
Fla..............	x	x	x	x	x			x	x				erning the apportionment of state
Ga..............	x	x	x	x	x		x		x	x			school funds (25)
Idaho..............	x	x	x	x	x	x	x	x	x				5. Regulation of teacher certification
Ill..............	No State Board												(37)
Ind..............		x	x	x	x					x			6. Regulation of teacher education
Iowa..............	No State Board												other than by certification (23)
Kans..............	x	x	x		x			x	x				7. Determination of the plan of or-
Ky..............	x	x	x	x	x		x	x					ganization for the state department
La..............	x	x	x	x	x		x		x				of education
Maine..............	No State Board												8. Adoption of courses of study (31)
Md..............	x	x	x		x	x	x	x		x	x		9. Adoption of Textbooks (21)
Mass..............	x	x			x	x	x						10. General control of state library
Mich..............		x			x	x							service (8)
Minn..............	x	x	x	x	x		x	x					11. Regulation of licensing in fields other
Miss..............		x		x	x		x						than education (2)
Mo..............	x	x	x		x	x	x	x					12. Management of state retirement
Mont..............	x	x	x		x	x	x	x					system for teachers (3)
Nebr..............	No State Board												
Nev..............	x	x	x		x		x	x	x			x	
N. H..............	x	x	x	x	x	x	x						
N. J..............	x	x	x		x	x	x	x		x			
N. Mex..............	x	x	x		x			x	x				
N. Y..............	x	x	x	x	x	x	x	x		x	x		
N. C..............	x	x	x	x		x		x					
N. Dak..............	No State Board												
Ohio..............	No State Board												
Okla..............	x	x	x	x	x			x					
Oreg..............		x	x	x	x	x		x	x				
Pa..............	x	x	x	x	x	x							
R. I..............	No State Board												
S. C..............	x	x	x		x			x	x				
S. Dak..............	No State Board												
Tenn..............	x	x	x	x		x		x	x	x			
Texas..............		x						x					
Utah..............	x	x	x	x	x		x						
Vt..............	x	x	x	x	x	x	x	x					
Va..............	x	x	x	x	x	x	x	x	x	x			
Wash..............	x	x	x	x	x	x		x					
W. Va..............	x	x	x		x	x		x	x				
Wis..............	No State Board												
Wyo..............	x	x	x	x	x	x	x					x	
No. states assigning...	32	37	34	25	37	23	21	31	21	8	2	3	

[a] Refers here only to boards responsible for elementary and secondary education.

[b] Numbers in parentheses show the number of states in which state boards of education have been assigned the functions described.

Source: *The Forty-Eight State School Systems* (Chicago: Council of State Governments, 1949), Appendix, Table 10, p. 184.

In order for the material to be scalable, the pattern of positive and negative responses in Table 18 must be of a special sort. Ideally, all the responses down to a certain point would be positive, and beyond this point all the responses would be negative. Only item number 3 fits this

Table 18. FIRST TRIAL ORDERING OF RESPONSES FOR SCALING

State	11 Yes	11 No	12 Yes	12 No	10 Yes	10 No	7 Yes	7 No	9 Yes	9 No	6 Yes	6 No	4 Yes	4 No	8 Yes	8 No	1 Yes	1 No	3 Yes	3 No	2 Yes	2 No	5 Yes	5 No
Calif......(11)		x	x		x		x		x		x		x		x		x		x		x		x	
N. Y......(10)	x		x		x			x	x			x	x		x		x		x		x		x	
Va.........(10)		x	x		x		x		x		x		x		x		x		x		x		x	
Ala........ (9)		x	x			x	x		x		x		x		x		x		x		x		x	
Ark....... (9)		x	x			x	x		x		x		x		x		x		x		x		x	
Idaho..... (9)		x	x			x	x		x		x		x		x		x		x		x		x	
Md........ (9)	x		x		x		x		x			x	x		x		x		x		x		x	
Wyo....... (9)		x	x	x		x		x	x			x	x		x		x		x		x		x	
Conn...... (8)		x	x		x			x	x		x		x			x	x		x		x		x	
Del........ (8)		x	x		x			x	x		x		x			x	x		x		x		x	
Ga........ (8)		x	x		x	x		x	x		x			x		x	x		x		x		x	
La......... (8)		x	x		x		x	x		x	x			x		x	x		x		x		x	
Nev....... (8)		x	x	x		x		x	x		x			x		x	x		x		x		x	
N. J....... (8)		x		x	x		x			x	x			x		x	x		x		x		x	
Tenn...... (8)		x		x	x		x	x		x	x		x			x	x		x		x			x
Vt......... (8)		x	x		x		x	x		x	x		x			x	x		x		x		x	
Fla........ (7)		x	x		x		x	x		x		x	x			x	x		x		x		x	
Ky......... (7)		x	x		x	x		x		x	x		x			x	x		x		x		x	
Minn...... (7)		x	x		x	x		x		x	x		x			x	x		x		x		x	
Mo......... (7)		x	x		x	x			x	x	x			x	x		x		x		x		x	
Mont...... (7)		x	x		x	x			x	x	x			x	x		x		x		x		x	
N. H...... (7)		x	x		x	x			x	x	x		x			x	x		x		x		x	
N. C...... (7)		x	x		x		x		x	x		x	x			x	x		x		x		x	
Oreg...... (7)		x	x		x		x		x	x		x	x			x	x			x	x		x	
Wash...... (7)		x	x		x		x		x	x	x		x			x	x		x		x		x	
W. Va..... (7)		x	x		x		x	x		x	x			x	x		x		x		x		x	
Kans...... (6)		x	x		x		x	x		x			x	x	x		x		x		x		x	
N. Mex.... (6)		x	x		x		x	x		x			x	x	x		x		x		x		x	
Okla...... (6)		x	x		x		x	x			x	x		x	x		x		x		x		x	
Pa......... (6)		x	x		x		x	x		x	x			x		x	x		x		x		x	
S. C....... (6)		x	x		x		x	x			x	x		x	x		x		x		x		x	
Utah...... (6)		x	x		x	x		x		x			x	x		x	x		x		x		x	
Ariz....... (5)		x	x		x		x	x		x		x	x		x		x		x		x		x	
Ind........ (5)		x	x		x		x	x		x		x	x		x		x		x		x		x	
Mass...... (5)		x	x		x	x		x		x	x			x	x		x		x		x		x	
Colo....... (4)		x	x		x	x		x		x	x		x	x		x		x		x	x			x
Miss...... (4)		x	x		x		x		x	x		x	x		x		x		x	x			x	
Mich...... (3)		x	x		x		x		x	x	x			x	x		x		x		x		x	
Texas..... (2)	x		x		x		x		x	x		x	x		x		x		x	x		x		x
		.2		3		8		21		21		23		25		31		32		34		37		37

Source: Table 17.

ideal pattern, though several others approximate the pattern.

When any of the items have more than two possible responses, it is frequently necessary to combine some of the response categories in order to approximate this pattern. In general, when it is necessary to combine some response categories in order to get a scale pattern, it is reasonable to assume that the combined categories were not separately useful in dif-

ferentiating the units being scaled along the dimension being measured by the scale and that for the purposes of this scale these categories should not have been made separate to start with.

When it is necessary to combine categories of multiple-choice items, the resultant categories are scored anew with integral values beginning with zero for the most negative response in each item. New total scores are then computed for each individual, and the individuals are re-ordered according to their new total scores.

Once the combination of categories is completed and the individuals are ordered according to their new total scores, we are at the point equivalent to Table 18. Next we rearrange the columns and rows of Table 18 in order to get the "best" scale pattern. Frequently several trials are necessary, the pattern being improved each time. Table 19 is our first rearrangement of Table 18. In Table 19 the states have been ordered so as to reduce the errors on either side of the tentative "cutting points" shown in Table 18. The states are arranged so that we get solid "runs" of X's in each column, insofar as possible.

An additional criterion is that no column should have more error than nonerror. The negative column of item 4 does not fulfill this criterion since it has eight errors and six nonerrors. If this were a multiple-response item, we would combine this category with another response category in an effort to improve the situation. When the item is dichotomous or has been reduced to a dichotomy, however, this is not a possibility. Since it does not seem possible to rearrange the states in such a way that this criterion can be met, we will omit item 4 as not scaling in the universe of content under consideration. An examination of the original items in Table 17 shows that item 4 is the only one dealing directly with fiscal matters, so it seems reasonable that it should not scale with the others.

In attempting to get a better arrangement of individuals it is frequently helpful to recopy Table 19, first copying all columns having positive responses, then copying all columns having intermediate responses, then all columns having negative responses. With this arrangement in an ideal scale, the responses form a parallelogram if the rows and columns are properly arranged. Table 20 shows Table 19 arranged in this parallelogram form with no re-ordering of individuals but with the columns arranged in order of cutting points. From Table 20 some of the immediately obvious changes to reduce error are as follows:

Move Colorado down 2 lines
Move Florida down 9 lines
Move Wyoming down 12 lines
Move Pennsylvania up 2 lines

Table 21 shows the final arrangement of the responses after these and other similar changes have been made.

Table 19. SECOND TRIAL ORDERING OF RESPONSES FOR SCALING

| | 11 | | 12 | | 10 | | 7 | | 9 | | 6 | | 4 | | 8 | | 1 | | 3 | | 2 | | 5 | |
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Ye	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
N.Y....	x		x		x				x		x	x	x		x		x		x		x		x	
Md.....	x		x		x				x		x	x		x	x		x		x		x		x	
Calif...		x	x		x				x		x		x		x		x		x		x		x	
Wyo....		x	x	x	x			x	x		x	x	x		x		x		x		x		x	
Nev.....		x	x	x		x	x		x			x	x		x		x		x		x		x	
Ga......		x	x		x	x	x		x		x		x	x		x	x		x		x		x	
N.J.....		x	x		x	x	x		x		x	x		x	x		x		x		x		x	
Tenn....		x	x		x	x		x	x		x		x		x		x		x		x			x
Va......		x	x		x	x	x		x		x		x		x		x		x		x		x	
Ala.....		x	x		x		x	x		x		x		x		x		x		x		x		x
Ark.....		x	x		x		x	x		x		x		x		x		x		x		x		x
Idaho...		x	x		x		x	x		x		x		x		x		x		x		x		x
Conn. ..		x	x		x		x	x		x		x	x		x		x		x		x		x	
Del.....		x	x		x		x	x		x		x	x		x		x		x		x		x	
La......		x	x		x			x	x		x		x		x		x		x		x		x	
Fla.....		x	x		x			x	x			x	x		x		x		x		x		x	
Vt......		x	x		x		x	x			x	x		x		x		x		x		x		x
Ky......		x	x		x		x	x		x		x	x		x		x		x		x		x	
Minn....		x	x		x		x	x		x		x	x		x		x		x		x		x	
Mo......		x	x		x		x	x		x	x		x		x		x		x		x		x	
Mont....		x	x		x		x	x		x	x		x		x		x		x		x		x	
N.H ...		x	x		x		x	x		x	x		x			x	x		x		x		x	
Oreg....	x		x		x			x	x		x		x		x		x	x	x		x		x	
Wash....	x		x		x			x	x	x	x		x		x		x		x		x		x	
W.Va...	x		x		x			x	x	x	x		x	x	x		x		x		x		x	
N.C....		x	x		x			x	x	x	x		x		x		x		x		x		x	
Okla....		x	x		x			x	x		x	x	x		x		x		x		x		x	
Pa......		x	x		x			x	x		x	x		x	x	x	x		x		x		x	
Utah....		x	x		x	x		x	x		x		x	x	x	x	x		x		x		x	
Kans....		x	x		x			x	x	x	x		x	x	x		x		x		x		x	
N.Mex..		x	x		x			x	x	x	x		x	x	x		x		x		x		x	
S.C.....		x	x		x			x	x	x	x		x	x	x		x		x		x		x	
Ariz.....		x	x		x			x	x	x	x		x	x	x		x	x	x		x		x	
Ind.....		x	x		x			x	x	x	x		x	x		x	x	x	x		x		x	
Mass....		x	x		x		x	x		x	x		x		x		x	x	x		x		x	
Colo....		x	x	x	x			x	x		x		x	x	x			x	x		x	x	x	
Miss....		x	x		x			x	x		x	x		x	x			x	x		x	x	x	
Mich....		x	x		x			x	x	x	x		x		x		x		x	x	x		x	
Tex.....		x	x		x			x	x	x	x		x	x		x	x		x		x		x	

Having made the arrangement of Table 21, we compute a *coefficient of reproducibility*. The coefficient of reproducibility measures the extent to which our original criterion of scalability has been fulfilled. This criterion stated that the responses should be reproducible from the rank of the individuals. Ideally, we would expect the states in the first rank

Table 20. PARALLELOGRAM ARRANGEMENT OF RESPONSES FOR SCALE

	Positive responses											Negative responses										
	11	12	10	9	7	6	1	8	3	2	5	11′	12′	10′	9′	7′	6′	1′	8′	3′	2′	5′
N. Y.	x		x		x	x	x	x	x	x	x		x		x							
Md.	x		x		x	x	x	x	x	x	x		x		x							
Calif.		x	x	x	x	x	x	x	x	x	x	x										
Wyo.		x			x	x	x	x	x	x	x	x		x	x							
Nev.		x		x	x		x	x	x	x	x	x		x			x					
Ga.			x	x	x		x	x	x	x	x	x	x					x	x			
N. J.			x	x	x	x	x	x	x	x	x	x	x	x					x			
Tenn.			x	x	x	x	x	x	x	x		x	x			x						x
Va.				x	x	x	x	x	x	x	x	x	x									
Ala.				x	x	x	x	x	x	x	x	x	x	x								
Ark.				x	x	x	x	x	x	x	x	x	x	x								
Idaho				x	x	x	x	x	x	x	x	x	x	x								
Conn.				x	x	x	x	x	x	x	x	x	x	x								
Del.				x	x	x	x	x	x	x	x	x	x	x								
La.				x		x	x	x	x	x	x	x	x	x		x						
Fla.				x		x	x	x	x	x	x	x	x	x		x	x					
Vt.				x	x	x	x	x	x	x	x	x	x	x	x							
Ky.				x		x	x	x	x	x	x	x	x	x	x			x				
Minn.				x		x	x	x	x	x	x	x	x	x	x		x					
Mo.				x	x	x	x	x	x	x	x	x	x	x	x							
Mont.				x	x	x	x	x	x	x	x	x	x	x	x							
N. H.				x	x	x	x		x	x	x	x	x	x	x				x			
Oreg.				x	x			x	x	x	x	x	x	x		x		x				
Wash.					x	x	x	x	x	x	x	x	x	x	x	x						
W. Va.				x	x	x	x	x	x	x	x	x	x			x						
N. C.				x			x	x	x	x	x	x	x			x	x					
Okla.							x	x	x	x	x	x	x	x	x	x	x					
Pa.					x	x	x	x	x	x	x	x	x	x	x				x			
Utah				x	x		x	x	x	x	x	x	x	x	x		x		x			
Kans.			x				x	x	x	x	x	x	x			x	x					
N. Mex.			x				x	x	x	x	x	x	x			x	x					
S. C.			x					x	x	x	x	x	x			x	x					
Ariz.				x			x	x	x	x	x	x	x			x	x	x				
Ind.				x				x	x	x	x	x	x			x	x	x	x			
Mass.				x	x	x		x	x	x		x	x			x			x	x		
Colo.		x				x			x	x		x	x		x	x	x	x	x	x	x	
Miss.						x		x	x	x	x	x	x	x	x	x	x	x	x	x		
Mich.				x				x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Tex.			x									x	x	x		x	x	x	x	x	x	x

Source: Table 19.

grouping, New York and Maryland, to have positive responses to all 12 items. Accordingly, New York has two errors and Maryland two. To compute the index of reproducibility, we first count the total number of

Table 21. FINAL ARRANGEMENT OF RESPONSES FOR SCALE MEASURING LEGAL AUTHORITY OF STATE BOARDS OF EDUCATION (SEE TABLE 16 FOR IDENTIFICATION OF ITEMS)

State	Scale group	11	12	10	9	7	6	1	8	3	2	5	11'	12'	10'	9'	7'	6'	1'	8'	3'	2'	5'
N.Y.	1	x		x		x	x	x	x	x	x	x		x		x							
Md.		x		x		x	x	x	x	x	x	x		x		x							
Calif.	2		x	x	x	x	x	x	x	x	x	x	x					x					
Nev.			x		x	x		x	x	x	x	x	x		x			x					
Ga.	3			x	x	x		x		x	x	x	x	x				x		x			
N.J.				x		x	x	x	x	x	x	x	x	x	x								
Tenn.				x	x		x	x	x	x	x	x	x	x			x						x
Va.				x	x	x	x	x	x	x	x	x	x	x									
Ala.	4				x	x	x	x	x	x	x	x	x	x	x								
Ark.					x	x	x	x	x	x	x	x	x	x	x								
Idaho					x	x	x	x	x	x	x	x	x	x	x								
Conn.					x	x	x	x	x	x	x	x	x	x	x								
Del.					x	x	x	x	x	x	x	x	x	x	x								
La.					x		x	x	x	x	x	x	x	x		x							
Wyo.	5		x			x	x	x	x	x	x	x	x		x	x							
Vt.					x	x	x	x	x	x	x	x	x	x	x	x							
Ky.					x			x	x	x	x	x	x	x	x	x		x					
Minn.					x			x	x	x	x	x	x	x	x	x		x					
Mo.					x	x	x	x	x	x	x	x	x	x	x								
Mont.					x	x	x	x	x	x	x	x	x	x	x								
N.H.						x	x	x		x	x	x	x	x	x					x			
Mass.						x	x	x		x	x	x	x	x	x					x	x		
Oreg.	6				x		x		x	x	x	x	x	x	x		x		x				
Wash.							x	x	x	x	x	x	x	x	x	x	x						
Pa.							x	x		x	x	x	x	x	x	x				x			
W.Va.					x		x	x	x	x	x	x	x	x		x							
Fla.	7				x			x	x	x	x	x	x	x			x	x					
N.C.					x			x	x	x	x	x	x	x			x	x					
Okla.								x	x	x	x	x	x	x	x		x	x					
Kans.					x			x	x	x	x	x	x	x			x	x					
N.M.					x			x	x	x	x	x	x	x					x	x			
S.C.					x			x	x	x	x	x	x	x			x	x					
Ariz.	8				x				x	x	x	x	x	x	x			x	x	x			
Ind.	9				x				x	x	x	x	x	x	x			x	x	x	x		
Utah						x		x	x	x	x	x	x	x	x	x		x		x			
Miss.	10								x		x	x	x	x	x	x	x	x	x		x		
Mich.							x		x		x	x	x	x	x	x	x	x	x	x	x		
Colo.	11			x				x			x	x	x				x	x	x		x	x	
Tex.	12				x						x	x	x				x	x	x	x	x	x	x

Source: Table 20.

errors. We then substitute in the following formula:

$$\text{Coefficient of reproducibility} = 1 - \frac{\text{number of errors}}{\text{number of questions} \times \text{number of individuals}} \qquad (1)$$

In Table 21 there are 36 errors of response.[9] Substituting in formula (33) we have:

$$\text{Coefficient of reproducibility} = 1 - \frac{36}{(11)(39)} = 1 - \frac{36}{429} = 1 - .084 = .916$$

It has been arbitrarily assumed that a coefficient of reproducibility of at least .90 is necessary before an area can be considered scalable.[10] Patterns of responses that can be made to take on a general rectangular distribution but whose coefficient of reproducibility is below .90 are said to form quasi-scales.[11] In our example, however, the coefficient of reproducibility is above the required level. Therefore, we can say, in accordance with prevalent conventions, that we have a scalable area.

When we look at the content of this scaled area, it seems reasonable to say that we are measuring the legal authority of state boards of education in matters other than financial. Such a measure would be useful in studies of centralization of authority, of development of educational systems, etc. Such a scale utilizes practically all the information available in the eleven items used.

Additional comments on the Guttman scale. When a Guttman scale is constructed in an attitudinal area and the items represent degrees of favorableness or unfavorableness, it is possible to determine an objective zero point on the scale which will divide the individuals into the "for" and "against" groups. This zero dividing point is relatively independent of bias in wording of the questions as long as the questions cover a reasonable portion of the range of attitudes.[12] Another important characteristic of a Guttman scale (which can be appreciated only after familiarity with Part IV of the text) is that the correlation between the scale scores and an outside measure (external to the scale) is the same as the multiple correlation between the outside measure and all of the items in the scale.

Another illustration of the Guttman scale. Gilbert Shapiro has utilized the Guttman scaling technique to rank states by degree of institutionalized segregation and discrimination.[13] The results are shown in Table 22. Shapiro comments on this as follows:

The scale, using the data from Myrdal's table, ranks the 23 states and the District of Columbia meaningfully from high to low according to the degree of institutionalized discrimination and segregation. Every state ranked higher than another in scale type uses all of the institutional modes of discrimination and segregation that the lower state uses, as well as at least one other.

[9] In this case the additional labor involved in making Tables 19 and 20 has hardly been an efficient expenditure of effort since Table 18 has only 41 errors when we omit item 4.

[10] See Stouffer et al., *Measurement and Prediction, op. cit.*, p. 77

[11] *Ibid.*, pp. 159–163.

[12] *Ibid.*, Ch. 7.

[13] Gilbert Shapiro, "Myrdal's Definition of the 'South': A Methodological Footnote," *American Sociological Review*, 13 (October 1948), pp. 619–621.

Table 22. RANKING OF STATES BY DEGREE OF INSTITUTIONALIZED SEGREGATION AND DISCRIMINATION

State	Entire State						Part of State			Absent from State						Scale Type
	A	B	C	D	E	F	A	C	F	A	B	C	D	E	F	
1. La.	x	x	x	x	x	x										
2. Ark.	x	x	x	x	x	x										
3. Miss.	x	x	x	x	x	x										
4. Ala.	x	x	x	x	x	x										I
5. Fla.	x	x	x	x	x	x										
6. Ga.	x	x	x	x	x	x										
7. S. C.	x	x	x	x	x	x										
8. Va.	x	x	x	x	x	x										
9. Tex.		x	x	x	x	x	x									
10. Tenn.		x	x	x	x	x	x									II
11. N. C.		x	x	x	x	x	x									
12. Okla.		x	x	x	x	x				x						III
13. Md.			x	x	x	x				x	x					IV
14. Ky.			x	x	x	x				x	x					
15. Del.				x	x	x		x		x	x					V
16. W. Va.				x	x	x				x	x	x				
17. Mo.				x	x	x				x	x	x				VI
18. D. C.				x		x				x	x	x		x		
19. Ind.					x				x	x	x	x	x			VII
20. N. J.										x	x	x	x	x	x	
21. Kan.										x	x	x	x	x	x	
22. Ill.										x	x	x	x	x	x	VIII
23. Ohio										x	x	x	x	x	x	
24. Pa.										x	x	x	x	x	x	

Key

A—White Primary
B—Jim Crow Street Cars
C—Jim Crow Railways
D—School Segregation Laws
E—Prohibition of Intermarriage
F—Actual Forced School Segregation

The District of Columbia has the sole error of reproducibility in the scale, lacking item E, prohibition of intermarriage, which is expected to be present in Scale Type VI.

Source: Gilbert Shapiro, "Myrdal's Definition of the 'South': A Methodological Footnote," *American Sociological Review*, 13 (October 1948), p. 621.

. . . Some of the data, which are taken from Myrdal, are now out-dated and the high degree of reproducibility of the scale (99.3 percent) indicates that it would probably be possible to add many more items characteristic of "south-

ernism," achieving more discrimination between states within the present scale types while still retaining a single continuum.

With the component dimensions of the concept "south" thus isolated, it is possible to observe relationships between these dimensions. Comparing Myrdal's historical characteristics with the scaled institutional dimension, we find that only those states in scale types I and II seceded during the Civil War, and only those in scale types I through VI were slave states or territories prior to 1860. The association between the institution scale and these historical events is perfect.[14]

Such a scale as this might also suggest the direction that an action program should take. If the purpose of a program is to move a state to a higher scale type (scale type VIII has less institutionalized segregation than scale type I), it would seem a logical hypothesis that one should attempt to change the items in the order that they change in going from lower to higher scale types.

CONSTRUCTION OF ARBITRARY INDEXES

Definitions. An arbitrary index is a composite index in which the relative importance of the individual items is decided arbitrarily rather than on some objective basis. Arbitrary indexes are of two types, weighted indexes and "unweighted" or "equally weighted" indexes. In an unweighted index the individual indexes are considered to be of equal importance, while in a weighted index some of the indexes are considered to be of more importance than others and are weighted accordingly.

When to use an arbitrary index rather than a scale.[15] In the first illustration used for the Guttman scale it would have been possible to construct an arbitrary index of legal authority by merely counting the number of areas in which each school board had authority and using this count as an index of legal authority. This would have assigned equal importance to each area of authority. The ranking of states by such an arbitrary method would be very much the same as the ordering arrived at by the Guttman scale. We are then faced with the question of why ever use a Guttman scale rather than an arbitrary index. If an area is scalable in the Guttman sense, the results from the scale will be very similar to the results from an equally weighted arbitrary index. However, unless the area is tested for scalability, we will not know whether our items are measuring more than one dimension or not. The process of scaling might show us that the area we are studying should be subdivided into several areas rather than studied as one variable that can be measured with a

---

[14] *Loc. cit.*
[15] For a more detailed discussion of this and related topics see Stouffer *et., al op. cit.* Chap. 6.

single score. For this reason it is advisable to test the scalability of an area before proceeding with an arbitrary index. Louis Guttman says,

Omitting a scale analysis and just going ahead with a single arbitrary index can completely obfuscate the purpose of the research, whether for descriptive or predictive purposes, if in reality several scores are required and not just one.[16]

He also says,

The existence or nonexistence of a scale is not a criterion of the worth of a problem. If a problem turns out to involve nonscalable data, they should be treated, not as scales, but in whatever manner will yield a proper answer.

To summarize briefly, problems which do not involve samples from a universe of items are not in general scale problems. Problems which do involve such sampling—including almost all of attitude and public opinion work—will profitably be studied first by means of scale analysis. If the universe is scalable, or can be broken down into scalable subuniverses, then it can be handled very easily by simple scale scores. How best to handle nonscalable universes remains a far more complicated, and as yet unsolved, problem.[17]

Arbitrary indexes are one means of handling nonscalable universes. In some cases factor analysis (Chapter 26) offers a possible solution.

AN ARBITRARY INDEX TO MEASURE ADEQUACY OF SCHOOL FACILITIES

The problem. If we were faced with the problem of rating several different school systems, perhaps the separate county systems in a state, we might construct an arbitrary index to do the job. A better solution might be to attempt to develop Guttman scales to measure the various aspects of adequacy in school systems or to develop a scale by the methods of factor analysis, treated in Chapter 26. However, if time and other resources were at a minimum and we were willing to settle for relatively crude results, we might choose an arbitrary index for the job. The following index is to illustrate the method of arbitrary index construction and is not presented as a valid index of the adequacy of school facilities.

Choice of items. The first problem is to select the individual items which we shall combine into the final index. We are, of course, limited to the consideration of measures that are available on all the school systems that we are interested in. From this list of available measures we would select the ones that we considered most important and that represented what we considered to be the various aspects of adequacy in a school system. For this selection of items and the evaluation of their importance we might utilize a panel of judges made up of specialists in the field.

[16] *Ibid.*, p. 175.
[17] *Ibid.*, p. 173.

Without going through a subjective evaluation procedure let us assume that the following items were selected either by a specialist or a group of specialists as the items to make up our index.

1. Average daily attendance as proportion of school age population
2. Expenses per pupil in average daily attendance
3. Value of school property per pupil in average daily attendance
4. Percent of teachers having at least bachelor's degree

Table 23. INCIDENCE OF ITEMS INDICATING ADEQUACY OF SCHOOL SYSTEMS FOR UNITED STATES AND SELECTED STATES

Item	Units	Observed values				Relatives [a]			
		U. S.	Ala.	Calif.	N. C.	U. S.	Ala.	Calif.	N. C.
1. Average daily attendance as proportion of school age population...	Percent	71.7	73.0	90.2	76.6	100.0	101.8	125.8	106.9
2. Expenses per pupil in average attendance.......	Dollars	179	99	235	110	100.0	55.4	131.8	61.8
3. Value of school property per pupil in average daily attendance....	Dollars	401	121	549	225	100.0	30.2	136.9	50.1
4. Percent of teachers having at least bachelor's degree.......	Percent	59.1	47.7	77.0	75.5	100.0	80.7	131.5	127.7

a. Obtained by expressing the state figure for an item as a percentage of the United States figure for the same item.

Source: *The Forty-Eight State School Systems* (Chicago: Council of State Governments, 1949), Appendix Tables 4, 7, 8, and 29.

Combination of simple indexes. Table 23 shows the measure of each of these items for the total United States and for three states. It is obvious that we cannot combine these measures directly and come out with anything meaningful. Therefore, our first step is to convert these *absolute* measures to *relative* measures. We do this by finding what percentage each measure is of the United States measure for that item. Thus, the value of school property per pupil in Alabama is only 30.2 percent of the value of school property per pupil in the United States as a whole. Table 23 also

shows these absolute values converted to *relatives*. These relatives we can combine into our final index.

If we decide that these four items are of equal importance in measuring the adequacy of a school system, then we combine them into an "unweighted" (equally weighted) index by simply averaging the relative values. For Alabama this gives

$$\frac{101.8 + 55.4 + 30.2 + 80.7}{4} = 67.0$$

We can symbolize this operation in a formula in the following way:

$$I = \frac{\sum \frac{p_n}{p_o} 100}{N} \qquad (2)$$

where I = the index

p_n = a measure of incidence of a trait for a state

p_o = a measure of incidence of a trait for the United States

N = the number of traits or indexes

With an index of this sort the index value for the United States will be 100, and the index values for the states will vary above and below this.

If we decide that these items are not of equal importance, we have to weight them in accordance with our judgment of their relative importance. Since the first item, average daily attendance as a proportion of school age population, does not include children attending private schools in some states, it is not as satisfactory an item for interstate comparisons as the others. For this reason we might consider it only one third as important as the other items for the purpose of constructing the index. We would weight each item accordingly (making the sum of the weights total 100 if feasible) and get a weighted average by dividing the sum of the weighted numbers by the sum of the weights. In this case for Alabama we would have

$$\frac{10(101.9) + 30(55.4) + 30(30.2) + 30(80.7)}{10 + 30 + 30 + 30} 100 = 60.1$$

We can symbolize this operation in the following formula:

$$I = \frac{\sum \frac{p_n w}{p_o}}{\Sigma w} 100 \qquad (3)$$

Where w is the weight assigned to each relative measure. If the student computes the index values for the states given by both the weighted and "unweighted" methods, he will find that the rank order is not changed by the change in methods. However, this is not necessarily the case.

Another illustration of an arbitrary index. In Robert Cooley Angell's monograph *The Moral Integration of American Cities* [18] several indexes are utilized. These are arbitrary indexes of various types and their construction is discussed in Appendix I of the monograph.[19] The quotation below gives the method of construction of one of these indexes, the Welfare Effort Index (Second Study).

The cards of Community Chests and Councils, Inc., show the population covered by the campaign in each city, the quota assigned, the amount raised, and the number of persons making pledges. From these data the following formula was constructed to give a welfare effort score:

$$\frac{\text{Amount raised}}{\text{Quota}} + \frac{\text{Pledgers}}{\begin{array}{c}\text{No. of families}\\ \text{in the area}\end{array}} + \frac{\text{Amount raised}}{\begin{array}{c}.0033 \times \text{Yearly}\\ \text{retail sales}\end{array}}$$

Each of the three ratios fluctuates around unity, so that the scores fluctuated around 3.0. It was thought that each of the three ratios measures one aspect of welfare effort—degree of achievement, proportion of families giving, and economic sacrifice involved.[20]

Other methods of constructing composite indexes. We shall mention two other methods commonly used by economists for combining simple indexes into a composite index, the method of "unweighted" averages of aggregates and the method of weighted averages of aggregates. In the former, instead of reducing each measure to a relative before combining them, one adds the several measures for a unit to get their "aggregate" and then divides this by the corresponding aggregate for the base year or area. In terms of the notation we have been using, the formula for the method of aggregates "unweighted" is as follows,

$$I = \frac{\Sigma p_n}{\Sigma p_o} \times 100 \tag{4}$$

Similarly the formula for the method of aggregates weighted is as follows:

$$I = \frac{\Sigma p_n q}{\Sigma p_o q} \times 100 \tag{5}$$

where p_n = measure of incidence of a trait for area n
     q = weight assigned to each measure of incidence

Since in the economist's indexes the p's usually refer to prices and the q's to quantities, their product is in units of dollars, no matter what the commodity involved. Therefore, adding these products seems logical. When

---

[18] Robert Cooley Angell, *The Moral Integration of American Cities* (Chicago: University of Chicago Press for *The American Journal of Sociology*, Vol. LVII, No. 1, Part 2, July 1951).
[19] *Ibid.*, pp. 123–126.
[20] *Ibid.*, pp. 124–125.

indexes are not in terms of money, however, the methods of combining relatives seem to be preferred over those of combining aggregates, which involve adding the values of such measures as number of radios and number of income tax returns.

The use of arbitrary indexes in sociology. With the development of the Guttman technique of scaling and other methods such as Paul Lazarsfeld's "latent attributes" [21] (not yet fully enough crystallized to be treated in a text of this type), arbitrary indexes are becoming of less importance in the field of sociology. Whenever some objective technique of weighting items can be utilized, the results are generally to be preferred to those obtained with weightings assigned from subjective judgment. The use of an "unweighted" index amounts to the assertion that the items are of equal importance. Unless the validity and reliability of an arbitrary index are thoroughly tested, the results are likely to be open to criticism.

Potentialities of indexes in sociological research. It seems that through the careful construction and standardization of indexes the sociologist will be able best to meet the challenge in the allegation that sociology can never become scientific because science demands measurement, while many of the phenomena of sociology are not the sort that can be counted or measured. By developing instruments for indirectly measuring these nonmeasurables, usually through measuring some of their more tangible correlatives, sociologists may be able to gain more precise and verifiable knowledge about them and their interrelations.

SUGGESTED READINGS

Chapin, F. S., *Experimental Designs in Sociological Research* (New York: Harper, 1947), Chap. 6.

Churchman, C. West, Ackoff, Russell L., and Wax, Murray (eds.), *Measurement of Consumer Interest* (Philadelphia: University of Pennsylvania Press, 1947).

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chaps. 20, 21.

Guttman, L. and Suchman, E. A., "Intensity and Zero Point for Attitude Analysis," *American Sociological Review*, 12 (February 1947), pp. 57–67.

Horst, P. and others, *The Prediction of Personal Adjustment* (New York: Social Science Research Council, 1941).

Merton, Robert K., and Lazarsfeld, Paul F. (eds.), *Continuities in Social Research: Studies in the Scope and Method of the "American Soldier"* (Glencoe, Ill., The Free Press, 1950).

Noland, E. W., "An Application of Scaling to an Industrial Problem," *American Sociological Review*, 10 (October 1945), pp. 631–642.

Sewell, William Hamilton, "Restandardization of a Sociometric Scale," *Social Forces*, 21 (March 1943), pp. 302–311.

Stouffer, Samuel A., and others, *Measurement and Prediction* (Princeton: Princeton University Press, 1950), Chaps. 1–9.

(Also see Suggested Readings of Chapter 26.)

---

[21] Stouffer *et al.*, *op. cit.*, Chaps. 10 and 11.

Time Series

THE PROBLEM, DATA, METHODS, AND USE OF TIME SERIES ANALYSIS

The problem. The basic fact of social change necessitates methods for studying variation of social phenomena in time. When a series of measures is recorded on some characteristic of a unit or a group of units for different periods or points of time, the record is called a time series. The methods for investigating the changes in time of the magnitudes of measures are called the methods of time series analysis. The sociologist's use of time series analysis in his attempt to understand the dynamic aspects of social phenomena is limited chiefly by the paucity of data relating to observations on the same type of phenomena over a long period of time. As societal record keeping improves, the use of the methods of time series analysis will probably become much more important in sociological research than at present.

Time series data available to the sociologist. An inventory of the more important sources of time series data available to the sociologist is fairly brief. First in importance is the series of decennial censuses, beginning in 1790. If all of the earlier censuses had secured data on the same subjects covered by the more recent censuses, and if all the data had been tabulated comparably, much research could be done in the analysis of changes that have taken place in the past 160 years with regard to various characteristics of the population. But for many of the characteristics sociologists are interested in, comparable data are not found at all or are found for only recent decennial years. Another handicap in the use of census material for time series analysis is the long period of time between observations. Economic data on production and prices are often available by single years, and often by months or even by days, so that the changes can be traced continuously. But even for the items on which successive censuses offer comparable data, there is a gap of 10 years during which the measures of the characteristic must be inferred from its values at the

beginning and end of the period. The five-year intervals between the censuses of agriculture since 1920 and the two-year intervals between the census of manufactures since 1919, until the interruption caused by World War II, are more satisfactory in this respect, but the information they offer of sociological interest is somewhat limited as is the length of time for which these more frequent censuses are available. Since the Census Bureau in 1942 took over from the Work Projects Administration the operation of a monthly sample survey originally designed for measuring unemployment, the scope of the information obtained has widened, and many types of data useful in social research are now available on a monthly or yearly basis from these current population surveys.

The annual sources of data on births and deaths issued by the National Office of Vital Statistics of the Public Health Service of the Federal Security Agency [1] are useful to sociologists. Some of the information is even recorded by months as well as by years. But it is only since 1933 that all the states have been in the birth and death registration areas, while the compilation of such data from the first group of states in the areas began only in 1915. Another great drawback to the use of the vital statistics for small geographic units is that, except in census years, we do not have a count of the corresponding population to which the births and deaths are related.

Other sources of time series data are given in the bibliography at the end of Chapter 3.

Development of the methods of time series analysis. As with the methods of index number construction, it is the economists who have developed most highly the methods of time series analysis. They apply these methods to direct measures and to indexes, both single and composite, computed for successive periods or points of time. The biometricians also have contributed methods, particularly in the fitting of growth curves to time series data. We shall sketch briefly the scope of the economists' methods of time series analysis and then present in more detail the parts of such methods that are applicable to sociological problems.

The economist's analysis of time series. [2] The economist in analyzing a time series tries to segregate the observed changes due to the four most important types of movements which occur in time: secular trend, cyclical,

---

[1] The volumes were called *Birth, Stillbirth and Infant Mortality Statistics* through 1936 when their names were changed to *Vital Statistics of the United States*, Part I, "Place of Occurrence" and Part II, "Place of Residence."

[2] The following paragraph is a condensation of parts of chapter 14, "The Problem of Time Series," in Frederick E. Croxton and Dudley J. Cowden, *Applied General Statistics* (New York: Prentice-Hall, 1939), pp. 363–384. The authors and publisher have kindly permitted reproduction of Charts 143A and 143B which are shown here as Figures 18 and 19. For the best introductory treatment available on time series analysis, the reader is referred to chapters 14–19 of *Applied General Statistics*.
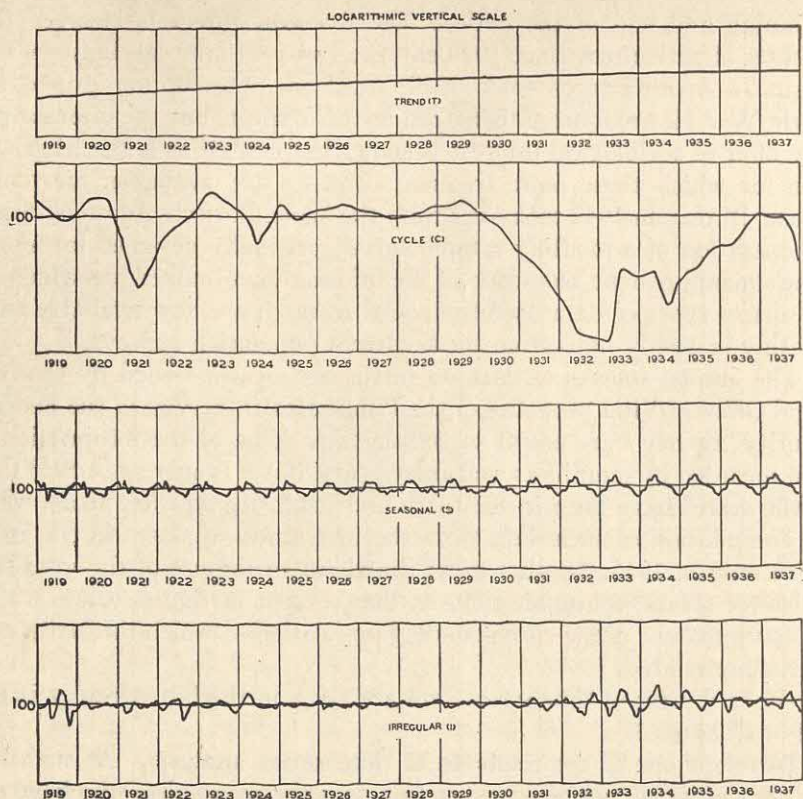
Figure 18.   Graphic Analysis of Variations in Pig Iron Production
in the United States, 1919–1937. Copyright by Prentice-Hall, Inc.

periodic, and irregular movements. Secular trend (T) refers to the long
time change. Cyclical (C) movements refer to those fluctuations associated
with "business cycles" which seem to occur about every four years in
general business but are not absolutely regular. Periodic variations are
those which recur at regular intervals of time within a year, month, day,
or hour; the periodic variations within the period of a year are called
seasonal variation (S). Irregular (I) or residual variations are all those not
included in the above three types; they may be either minor random
movements, or episodic movements. Figures 18 and 19 portray these
types of movements graphically. The series in Figure 18 show the four
component types of movements which have been isolated by analysis.
They are plotted on a logarithmic vertical scale so that graphic addition
of their magnitudes is the equivalent of multiplication. Figure 19 shows
the synthesis of the same components in the successive combinations,
T, $T \times C$, $T \times C \times S$, and $T \times C \times S \times I$. The last one is the record
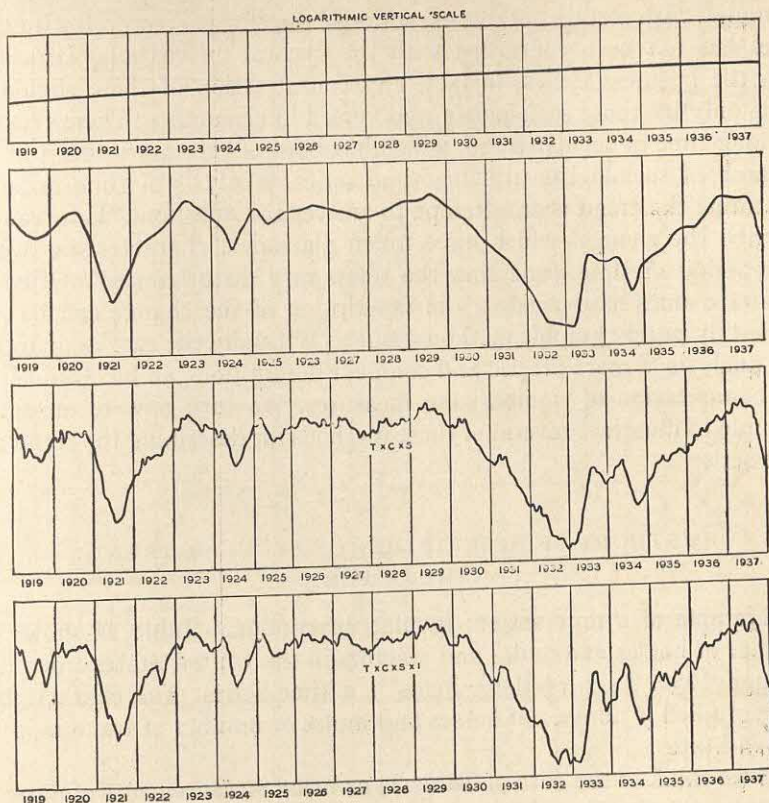
Figure 19. Graphic Synthesis of Variations in Pig Iron Production in the United States, 1919–1937. Copyright by Pentice-Hall, Inc.

of the original observations from which the components have been analyzed by various methods. Sometimes interest may be focused upon the trend, T, to study long time changes. Often with economists the interest is in the cyclical movements, C, either alone or combined with the trend. For special purposes seasonal fluctuations, S, may be the primary object of interest. Although economists are by no means unanimous in agreement on all of their methods of analysis and especially on the interpretations of the results, they have a rather well developed body of method and theory relating to time series analysis.

Type of time series analysis used by sociologists. Only a part of the analysis outlined above is appropriate in the investigation of sociological time series. Because of the limitations of sources, few data are available on sociological phenomena by shorter periods of time than years, and therefore seasonal fluctuation is not often a factor to be considered. Fur-

thermore, with certain notable exceptions,[3] sociological research with time series has not been concerned with the cyclical fluctuations associated with the business cycles; in fact, any consideration of these should be made only by those sociologists well versed in economics. There remains the long-time or secular trend, which has been of primary interest in the research of sociologists utilizing time series. Analysis of time series to determine the trend is an attempt to answer the question, "How can we describe the changes which have taken place in a characteristic over a long period of time neglecting the temporary disturbances?" Often no elaborate analysis is needed; the description of the change can be performed by purely graphic methods. But as is usually the case in quantitative analysis, a more precise and compact description can be obtained by the computation of summarizing measures. We turn now to an actual example to illustrate several of these methods for describing the trend of a time series.

METHODS OF DESCRIPTION OF SECULAR TREND: LINEAR FORM

Example of a time series: tabular description. Table 24 shows the number of horses and mules and tractors in the United States from 1918 through 1950. Each of these items is a time series. The characteristic being studied is number of horses and mules or number of tractors in the United States.

If we think of the data of Table 24 as being measurements of a special sort taken in the United States each year, the major difference between a table showing time series data and a table showing a quantitative distribution, such as Table 8, becomes clear. In a table showing a quantitative distribution the data are measurements taken on a number of individual units at one time, while in a time series table the data are measurements on one individual unit taken at a number of times. (In Table 24 the individual unit is the United States.) It is important to recognize this fundamental distinction between the two. For even if the measures shown in the time series were treated as the ungrouped measures of a quantitative distribution, that is, if they were grouped into class intervals and condensed into a frequency table, they would constitute a different type of frequency distribution from the sort we have considered so far, since the unit whose measures vary is the same throughout and since the variation is therefore associated with differences in time rather than with differences in individuals.

---

3. The most important exceptions are Dorothy S. Thomas, *Social Aspects of the Business Cycle* (London: George Routledge, 1925), and *Social and Economic Aspects of Swedish Population Movements, 1750–1933* (New York: Macmillan, 1941).

Table 24. NUMBER OF HORSES AND MULES, AND TRACTORS ON FARMS IN THE UNITED STATES, JANUARY 1, 1918–1950

Year	Horses and mules (thousands)	Tractors (thousands)
1918	26,723	85
1919	26,490	158
1920	25,742	246
1921	25,137	343
1922	24,588	372
1923	24,018	428
1924	23,285	496
1925	22,569	549
1926	21,986	621
1927	21,192	693
1928	20,448	782
1929	19,744	827
1930	19,124	920
1931	18,468	997
1932	17,812	1,022
1933	17,337	1,019
1934	16,997	1,016
1935	16,683	1,048
1936	16,226	1,125
1937	15,802	1,230
1938	15,245	1,370
1939	14,792	1,445
1940	14,478	1,545
1941	14,104	1,675
1942	13,655	1,885
1943	13,231	2,100
1944	12,613	2,215
1945	11,950	2,422
1946	11,063	2,585
1947	10,021	2,800
1948	9,130	3,150
1949 [a]	8,246	3,500
1950 [a]	7,463	3,825

[a] Preliminary
Source: *Agricultural Outlook Charts* (Washington: Bureau of Agricultural Economics.)

Although this distinction must be kept in mind, many of the methods we have presented for the ordinary quantitative distribution can be used for description of this time series. For instance, if we want to know the average number of horses and mules on farms during this 33-year period, we simply evaluate formula (2) of Chapter 8,

$$\bar{X} = \frac{\Sigma X}{N}$$

thus,

$$\bar{X} = \frac{576,362,000}{33} = 17,466,000$$

If one were describing very briefly the number of horses and mules on farms in the United States during this period and wished to use only one measure, this mean number of 17,466,000 is probably the figure we would choose.

If we wish to go a step further in description and consider variation, however, we should not use the standard deviation. There are several reasons for this, some of them involving complicated matters such as the lack of independence between the observed measures, to be mentioned later in the chapter, but the major reason is obvious from an inspection of Table 24. The most conspicuous feature in the variation in the number of horses and mules is not the magnitude of the variation, which the standard deviation would describe, but the regular progressive decrease in number during the 33-year period. This phenomenon of progressive change in time is called secular trend, and it is the particular aspect of variation in number of horses and mules that we want to describe.

Graphic description. Graphic forms in the analysis and presentation of time series are an exceedingly important and useful method, either alone or as supplementary to some more elaborate method. Figure 20 shows the number of horses and mules and tractors on farms for each year from 1918 through 1950, with lines joining the plotted points. Let us differentiate between quantitative distributions and time series in the matter of constructing coordinate charts such as this. It will be remembered that grouping was necessary for the data of an ordinary quantitative distribution before such charts could be constructed (for presenting a distribution of measures on a single characteristic). But here, since the differentiating aspect of the several observations is not the identity of the individuals measured, but is *time*, which in itself is an ordered continuum, it can be represented along a graduated axis, whereas it would have been meaningless to measure off the serial numbers of the women in Table 8 on such an axis.

The chart shows very clearly the negative direction of change in num-
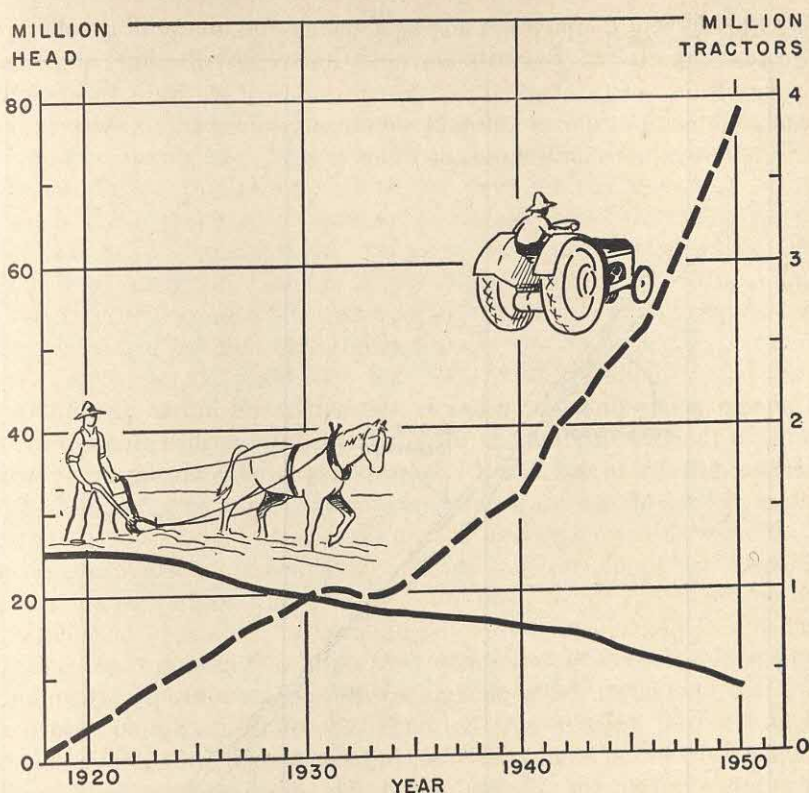
Figure 20.   Horses and Mules and Tractors on Farms, January 1, 1918, through 1950. (Source: Table 24.)

ber of horses and mules on farms and the positive direction of change in number of tractors. These facts are obvious even though there are some irregularities in number of tractors during the early thirties. It can be seen from Figure 20 that the decreasing trend in number of horses and mules is much closer to a straight line than is the increasing trend of tractors. For this reason we will confine most of the following discussion to number of horses and mules. We refer the reader to Figures 8 and 9 and the discussion of them in Chapter 5 for a treatment of the use of logarithmic vertical scales in charts of time series and in making graphic comparisons of two time series.

Description of trend by method of inspection.   To describe the long time trend more precisely, we wish to eliminate the irregularities resulting from disturbing influences, whatever they are, which give the line representing number of horses and mules in Figure 20 its irregular appearance. An easy way of doing this is simply to draw the straight line which
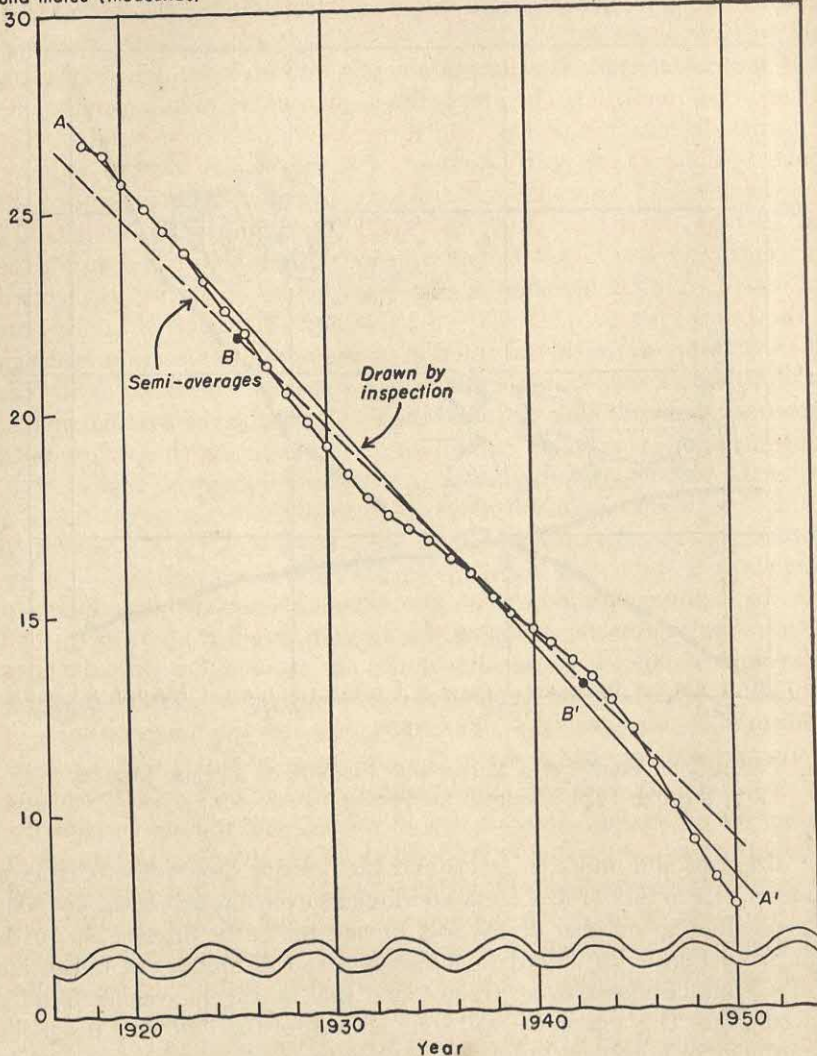
Figure 21. Horses and Mules on Farms, January 1, 1918, through 1950 with Straight Line Trends Fitted by Inspection and by the Method of Semi-Averages. (Source: Table 24.)

appears to approximate the broken line most closely. This is called fitting a straight line by the method of inspection. It is most conveniently accomplished with the use of a transparent ruler. Figure 21 shows the plotted points for horses and mules from Figure 20 with a straight line,

AA', fitted to them by inspection. This line AA' graphically describes the trend in numbers of horses and mules on farms during the period from 1918 to 1950.

If we wish to translate this graphic picture into a numerical measure, we need two quantities. The first is the mean number of horses and mules on farms during this period, which we have already computed. The second is the average yearly amount of change which we can estimate from the line AA'. Note that AA' crosses the vertical line representing the year 1918 at about the value of 26,800,000 and crosses the vertical line representing 1950 at about the value 8,250,000. Thus, we can compute the total decrease in the number of horses and mules on farms, as pictured by the trend line, as 26,800,000 − 8,250,000 or 18,550,000 in the 33-year period. If we perform the subtraction in the opposite direction, the sign of the difference will show the direction of change, and if we divide the difference by the number of years, the result will be the average annual change in number of horses and mules on farms during the period, thus,

$$\frac{8,250,000 - 26,800,000}{33} = -\frac{18,850,000}{33} = -571,200$$

With this figure computed we can now expand our textual description to the following statement: "During the 33-year period from 1918 to 1950 the average number of horses and mules on farms in the United States was 17,466,000, while the number decreased by about 571,200 a year."

Method of semi-averages. For many purposes the rough estimate of the average annual change secured above may be satisfactory, but its accuracy will vary with the ability and experience of the person fitting the line by inspection. Another simple way of determining the line describing the trend is the method of semi-averages. Here we take the mean number of horses and mules on farms for the first half of the period and plot this value opposite the midpoint of time for the first half period. Similarly we take the mean number of horses and mules for the second half of the period and plot it opposite the midpoint of time for the second half of the period. In our example the dividing of the 33-year period into halves is complicated by the fact that there is an odd number of years. The simplest way to get around this difficulty is to count the middle year, 1934, in both of the halves. The sum of the number of horses and mules for the first 17 years is 371,660,000, which divided by 17 gives 21,862,000 as the semi-average for the first half period. The sum of the last 17 figures is 221,699,000, which divided by 17 gives 13,041,000 as the semi-average for the second half period. Figure 21 shows these two values plotted opposite the points midway between 1925 and 1926 and midway between 1942 and 1943 respectively. The straight line labeled B and B'

is a line drawn through these two points. This line is called the line of semi-averages. It crosses the vertical line for 1918 at about 25,800,000 and the vertical line for 1950 at about 9,200,000. We can use these values to compute another estimate of the average annual rate of change thus,

$$\frac{9,200,000 - 25,800,000}{33} = \frac{-16,600,000}{33} = -503,000$$

This method of fitting a straight line by semi-averages has an advantage over the method of fitting by inspection in that every one who uses it will get the same results, whereas those who use the method of fitting by inspection may get somewhat different results.

Meaning of the straight lines fitted. Let us compare the meaning of the broken line joining the points representing the original observations in Figure 20 and the straight lines of Figure 21. The broken line of Figure 20 from left to right represents the actually observed number of horses and mules on farms during the period, with the differences in height between any two adjacent years representing the change in number of horses and mules from one year to the next. The time series is a composite of at least three types of movements—the movement of the long-time trend, cyclical movements, and irregular movements. We have no methods of treating the irregular movements systematically; we should have to employ the rather elaborate techniques of the economist to treat the cyclical movements; but we can approximate a description of the secular trend when other movements are eliminated from the observed data. This is exactly what lines AA' and BB' attempt to do—to show what the number of horses and mules would have been for each year during the period *if* cyclical and irregular influences had not been operating. We cannot be sure, however, that the cyclical and irregular movements do not affect our estimates of the trend because the only method we have used to eliminate them is to assume that their positive and negative variations will cancel one another out.

Assumptions involved in fitting a line to describe the trend. In fitting a line to the observed points and calling it the "trend" we are making two assumptions. First, we are assuming that there was some degree of regularity in the change going on, which we can try to separate from the observations which are composites of regular and irregular changes. That is, we are assuming that the concept of "trend" or "long-time change" is valid in this case. Does this mean then that we are assuming a "law" of change? Or that we could predict that the trend would be continued exactly the same for the next 33 years? Because of the orderliness and regularity of certain phenomena, it is possible that we might try to throw light on some of the above questions from analysis of the time series under investigation. But let us emphasize that this would be carrying the

problem over into the realm of inductive statistics. We mention the possibility of extensions of use of information about a particular unit in a particular period of time in order to emphasize that extensions of historical descriptions are possible, but that the validity of the historical description is not dependent on the validity of the extensions, which will be discussed later. What we are seeking now is a convenient description of the number of horses and mules on farms in the United States from 1918 to 1950. And because lines may be projected or interpreted as predictions in inductive statistics, they do not have to be so projected or interpreted in descriptive statistics. Therefore, the assumption made is only that there was between 1918 and 1950 a continuous decline in the number of horses and mules which can be approximately pictured by a line or curve of some sort.

The second assumption is that the trend explained above can be pictured by a straight line; that is, that the average annual amount of change was uniform throughout the entire period. The justification for this assumption is that on Figure 21 we can see that the straight lines appear to fit the observed points fairly closely. If we were discussing the changes in number of tractors on farms, however, it seems reasonable from Figure 20 that a more complex curve might describe the data better.

Method of moving averages. A straight line does not always best describe the trend exhibited by the observed points of a plotted time series. When this is the case, the method of moving averages can be used as a graphic description of the trend. If for each year we plot the average (arithmetic mean) of its own value and the values of the two adjacent years, a considerable amount of the irregularities are reduced or "smoothed." The line joining such points is known as a three-year moving average. There will be no three-year average point for the first year or the last year because there are no earlier or later data to average in with the other values.

A moving average for a longer period smooths the fluctuations more than one for a shorter period of time. For example, a seven-year moving average would be smoother than a three-year moving average. However, there would be no seven-year moving average points for the first three or the last three dates.

The economist often uses moving averages to smooth out cyclical movements so that the result will be a *trend* and not include cyclical movements also. In such a case the length of the period of the average should be the same as that of the cycle to be eliminated (so that it will include an equal amount of the higher and lower values of the cycle) or an integral multiple of that length. Since cycle lengths may be an even number of years, instead of an odd number like three or seven, it may be necessary at times to plot a two-year moving average or a four-year one.

When the number of years is even, the computation requires another step, which we call "centering" because the center of an even-year average falls between two years. After the even-year moving averages are computed, each adjacent pair of them is again averaged together, giving a point which falls opposite a year to be plotted.

The use of a moving average is indicated in simple analyses of sociological time series in several situations:

1. Where one wishes to eliminate cyclical fluctuations, in which case the period of the average should be chosen to be the same as that of the cycle or an integral multiple of it. (It is often desirable to examine the material in smoother form than that of the broken line of the observed data in order to decide whether it would be better to try to fit a straight line or a curve.)

2. Where a straight line does not fit the data, but where one does not wish to perform an analysis so elaborate as that of fitting a complex curve.

Summary of graphic methods of describing the trend. So far we have obtained or discussed approximate descriptions of the trend by the following methods, which may all be called graphic although some of them require a limited amount of preliminary arithmetic: (1) the broken line joining the points representing the observations; (2) a broken line joining the points representing a moving average; (3) the straight line fitted by inspection to the points representing the observations; (4) the straight line fitted by the method of semi-averages to the points representing the observations.

All of these methods require tabular or graphic forms for presentation. The next method we shall present may be presented by a table or chart, but it does not require either form. The method defines the line or curve by means of an algebraic equation, and because of its precision and compactness, it is the way a trend is usually described. The particular group of procedures we shall use for determining the equation in this example is called the method of "least squares," because the criterion used to determine the fit of the line or curve is that of minimizing certain squares. We shall illustrate the method of least squares for only the simplest case— that of fitting a straight-line trend.

Geometric concepts and conventions basic to describing a trend by an algebraic equation. The expressing of a geometric form such as a straight line by means of an algebraic equation involves the principles of analytic geometry. The derivation of the method involves differential calculus. And yet the determining of the equation for a particular case, the plotting of the equation on a chart, and even the interpreting of the constants of the equation involve no more mathematical ability than one should have acquired in high school algebra. Since many sociology students are quite a few years away from their high school algebra, however, we shall sum-
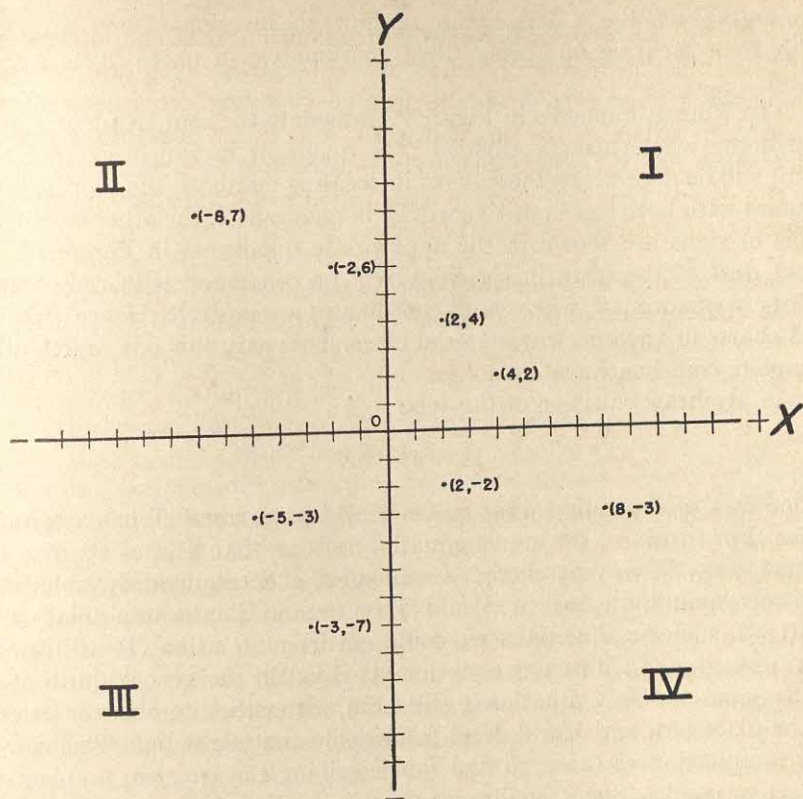
Figure 22. Location of Points in a Plane by Cartesian Co-ordinates.

marize a few elementary principles and conventions which one needs to have in mind before he proceeds with an analysis.

The fundamental principle of plane analytic geometry is that of a coordinate system, in which any two numbers locate a point with respect to two perpendicular graduated axes, a horizontal axis usually called the X-axis and a vertical axis usually called the Y-axis. It is conventional to set up the axes with their positive and negative directions as shown in Figure 22. The *abscissa*, the first of a pair of numbers locating a point, is used to indicate the number of units to be measured from the origin, 0, *along* or *parallel* to the X-axis, which is the same as the *perpendicular* distance from the Y-axis. The *ordinate*, the second number of the pair, is used to indicate the distance to be measured *along* or *parallel* to the Y-axis, which is the same as the *perpendicular* distance from the X-axis. For instance, the point (4, 2) is four units to the right of the Y-axis and

two units above the X-axis. Figure 22 shows the locations of several points with their locating numbers, called coordinates, in parentheses beside them.

The Roman numerals in Figure 22 designate the four quadrants into which the whole plane is divided. The quadrant in which a particular point will lie depends on the signs of its locating numbers, the coordinates. A point with both signs positive will lie in quadrant I; the other combinations of signs are shown in the appropriate quadrants in Figure 22. A great deal of the work in sociology will be concerned with only those points in quadrant I, where both coordinates are positive. Hence, graphs and charts used to present material often show only this one fourth of a complete coordinate system.

An algebraic equation of the type

$$Y = a + bX \tag{1}$$

is one way of expressing what value Y will have when X has a certain value. For instance, the above equation tells us that if $X = 0$, $Y = a$. Thus for any X we may choose, a value of Y is determined. If we let the two corresponding values of X and Y be the coordinates of a point on a chart, all such possible pairs we could get from equation (1) will lie on a straight line. In statistics equation (1) is called the general form of a linear equation.[4] Any equation in this form, with either numbers or letters in the place of a and b, will describe or define a straight line. This is the sort of equation we want to find for describing the trend in number of horses and mules. Since the form is prescribed, all we have to do is to find the appropriate numerical values for a and b.

As we have seen, several slightly different lines appear to fit the data equally well. But there is one line, which passes through the value of the mean number of horses and mules at the middle year, and which fulfills the criterion of "least squares"—that is, the sum of the squares of the vertical deviations of the observed points from the line is less than the corresponding sum of squares of deviations from any other line which can be drawn. Thus, the line may be considered as a sort of extension of a mean, which has the same property for one point in time. The sum of the squares of deviations of observed values is smaller when measured from the mean than when measured from any other value. And similarly the line from which the sum of the squares of the vertical deviations of the observed values is the smallest is the line called the "best" fit by the least squares criterion. There are other possible criteria which are sometimes used, such as the requirement that the sum of the squares of the perpendicular distances to the line be a minimum, or the requirement that

---

[4] In mathematics equation (1) is called the "slope-intercept" form of a linear equation.

the sum of the absolute values of vertical deviations be a minimum. However, the least squares criterion applied to the vertical deviations is the one used most frequently both in time series and, as we shall see, in regression analysis.

The problem is to determine the values of a and b which substituted in the equation,

$$Y = a + bX \tag{1}$$

will algebraically describe a line so that the sum of the vertical distances of the observed points from this line will be a minimum. Ordinary arithmetic or algebra cannot solve such a problem. Determination of the values which will yield a minimum requires differential calculus. Fortunately for those who do not have a mastery of calculus, however, the problem can be solved once, using letters instead of numbers, and the results expressed in simple algebraic formulas which can be evaluated for any particular example by one who has no knowledge of the calculus by which the formulas were derived.

Actual procedures in fitting a line by the method of least squares. First, we must assign letters to our variables so that we shall know what numerical values to put into the formulas. There are in our case two sets of varying values, the successive years which denote successive periods of time and the number of horses and mules observed for each of these years. Whenever time is one of the variables, we usually call the different years the X's, since the X-axis is assigned to that variable which we think of as independent, if there is one. Time pursues its course at the same rate year by year and is therefore regarded as independent. Instead of using the year 0 A.D. as the zero point on our X scale, however, we shall use the year 1918 as zero, purely for the sake of the convenience of working with small numbers. This makes 1919 have the value of 1 and so on up to 1950, which has the value of 32.

The characteristic, "number of horses and mules on farms," will be assigned to the Y-axis, which is usually reserved for the "dependent" variable. Now this use of "dependent" means merely that the number of horses and mules on farms is related in some way to the time at which this characteristic is observed. It does not imply that the mere passage of time *causes* changes in the number of horses and mules; the causes of change may be many and varied—but whatever they are, at least some of them are also associated with changes in time, and therefore the numbers of horses and mules for different years are different in value. The number of horses and mules in their chronological order will be the successive Y values. The formulas derived by the use of differential calculus for determining the values of a and b which will satisfy the least squares criterion for the vertical deviations are as follows:

Table 25. COMPUTATIONS FOR FITTING A STRAIGHT LINE BY THE METHOD OF LEAST SQUARES TO DATA FROM TABLE 24

Year (1)	X (1) $-$ 1918 (2)	Y Number of horses and mules (100,000's) (3)	XY (2) $\times$ (3) (4)	x (1) $-$ 1934 (5)	xY (3) $\times$ (5) (6)
1918	0	267	0	-16	-4272
1919	1	265	265	-15	-3975
1920	2	257	514	-14	-3598
1921	3	251	753	-13	-3263
1922	4	246	984	-12	-2952
1923	5	240	1200	-11	-2640
1924	6	233	1398	-10	-2330
1925	7	226	1582	-9	-2034
1926	8	220	1760	-8	-1760
1927	9	212	1908	-7	-1484
1928	10	204	2040	-6	-1224
1929	11	197	2167	-5	-985
1930	12	191	2292	-4	-764
1931	13	185	2405	-3	-555
1932	14	178	2492	-2	-356
1933	15	173	2595	-1	-173
1934	16	170	2720	0	0
1935	17	167	2839	1	167
1936	18	162	2916	2	324
1937	19	158	3002	3	474
1938	20	152	3040	4	608
1939	21	148	3108	5	740
1940	22	145	3190	6	870
1941	23	141	3243	7	987
1942	24	137	3288	8	1096
1943	25	132	3300	9	1188
1944	26	126	3276	10	1260
1945	27	120	3240	11	1320
1946	28	111	3108	12	1332
1947	29	100	2900	13	1300
1948	30	91	2730	14	1274
1949	31	82	2542	15	1230
1950	32	75	2400	16	1200
Sums:	528	5,762	75,197	0	$-16,995$

$\Sigma X^2 = 11,440$ $\Sigma x^2 = 2,992$

Formulas:

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y - b\Sigma X}{N}$$

$$Y_c = a + bX$$

Evaluations:

$$b = \frac{33(75,197) - (528)(5762)}{33(11,440) - (528)^2}$$

$$= -5.680$$

$$a = \frac{5762 - (-5.680)(528)}{33}$$

$$= 265.49$$

$$Y_c = 265.49 - 5.680X$$

Table 25. (CONTINUED)

Alternate method

$$b = \frac{\Sigma xY}{\Sigma x^2}$$

$$a' = \frac{\Sigma Y}{N}$$

$$a = a' - b\left(\frac{N-1}{2}\right)$$

$$Y_c = a + bX$$

$$b = \frac{-16,995}{2,992}$$

$$= -5.680$$

$$a' = \frac{5762}{33}$$

$$= 174.61$$

$$a = 174.61 + 5.680(16)$$

$$= 265.49$$

$$Y_c = 265.49 - 5.680X$$

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \tag{2}$$

and $\quad a = \dfrac{\Sigma Y - b\Sigma X}{N} \tag{3}$

where X = successive values of the independent variable

Y = observed values of the dependent variable

and N = number of observations

It will be noted that the formula for a has a b in it. The formula could be written wholly in terms of X and Y but this is the more condensed form. The above formula requires one first to find b and then to substitute this value in (3) to find a.

The computations required to get the quantities needed to evaluate equations (2) and (3) are not difficult. N is simply the number of observations made, which in this case is 33. Notice that N is one greater than the largest X since the first X value is zero. The two terms ΣX and ΣX^2 are simply the sum of the first 32 integers and the sum of the squares of the first 32 integers. These sums can be looked up in appendix tables of almost any statistics text in economics.[5] The other terms required are ΣY and ΣXY which can be calculated very easily on a calculating machine.[6] Table 25 shows these sums as well as the intermediate cross-products, the XY's. In practice it is unnecessary to compute these individual XY's. With the values shown as sums in the lowest row of the computation table, we evaluate formulas (2) and (3) and get

$$a = 265.49$$
$$b = -5.680$$

These values determine the equation of the straight line we want. However, if we substitute them directly into equation (1), we will get an

---

[5] For example, Croxton and Cowden, *op. cit.* Appendix M, p. 889.
[6] See Katharine Pease, *Machine Computation of Elementary Statistics* (New York: Chartwell House, 1949), Chap. 4.

equation which will describe the number of horses and mules in 100,000's since Y in Table 25 is given in 100,000's. Therefore, we multiply each of these values by 100,000 before substituting in equation (1). Performing this operation and substituting, we have

$$Y_c = 26,549,000 - 568,000X \tag{4}$$

Notice that we have added the subscript c to the Y in this equation for the purpose of differentiating a value computed from this equation, now denoted by Y_c, from an originally observed Y, written without subscript.

The computation table contains also an alternate method of computation where the x's, denoting deviations from the mean year, are used instead of the X's. It is somewhat shorter, since the equations for (1) and (2) are reduced when the x's are used. The substitutions are shown in Table 25. The values obtained for a and b are identical with those obtained by the first method.[7]

To draw the line described by equation (4), we have only to choose two values for X, substitute them one at a time in the equation, and solve each time for the corresponding value of Y_c. These two pairs of values will determine two points on a chart with the X-axis representing time with 1918 as zero and the Y-axis representing the number of horses and mules on farms. Since drawing a line is accomplished more accurately when the two points chosen are not too close together, let us choose for the X values the numbers representing the first and last years, 0 and 33. Substituting, we find,

when $X = 0, \quad Y_c = 26,549,000$
when $X = 33, Y_c = 7,805,000$

Figure 23 shows these two points plotted and the line which they determine as well as the originally observed points.

Let us examine the meaning of the constants a and b in equation (4). We have just seen that a is equal to the value of Y_c corresponding to a value of zero for X. This relation is true for every linear equation expressed in the form,

$$Y = a + bX \tag{1}$$

In trying to visualize geometrically the line described by an equation in this form without actually plotting the line, the first step is to imagine a point on the Y-axis located a units above the X-axis. This is the starting point of the line, always easy to locate since it has the coordinates $(0,a)$. We need only one more point to determine the line. The constant b tells

---

[7] For a modification of the alternate method to be used in the case where the number of observations is an even number, and there is no middle year, see Croxton and Cowden, *op. cit.*, pp. 405–407.
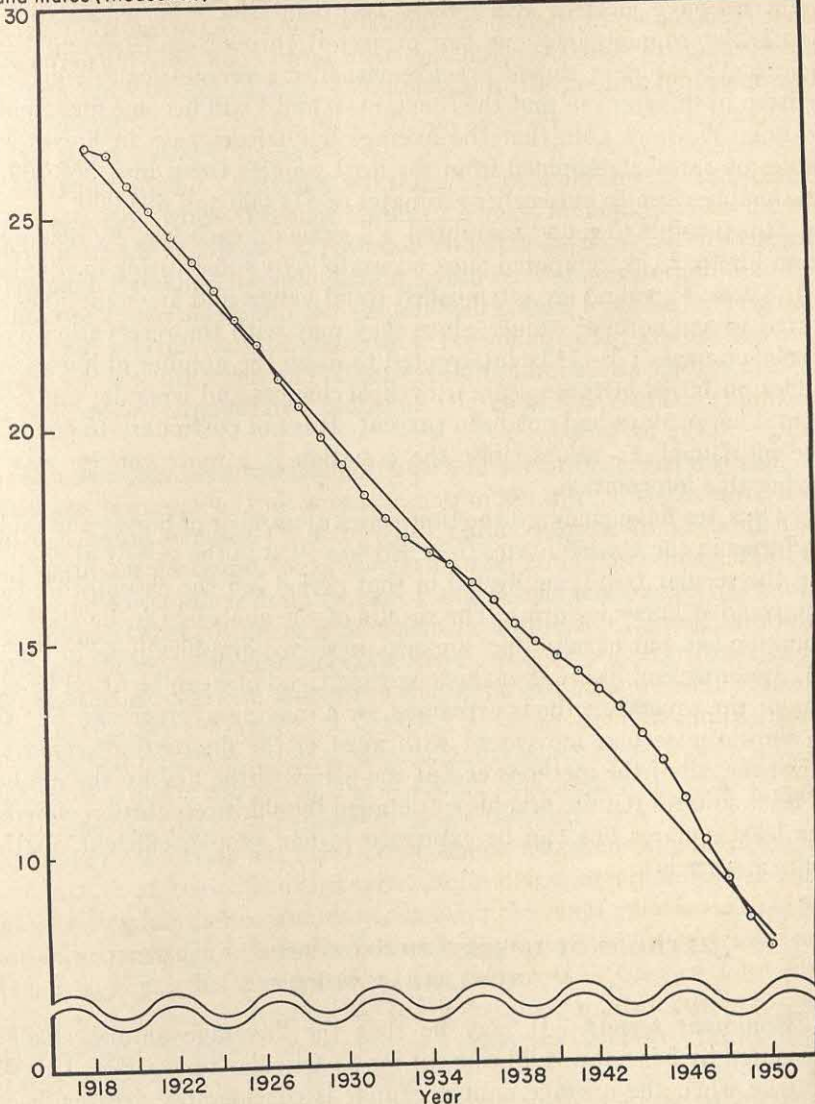
Figure 23.   Number of Horses and Mules on Farms in the United
States, 1918–1950, with Straight Line Trend Fitted by the Method
of Least Squares. (Source: Table 23.)

how many units of change occur in Y_c as X increases one unit, and the
sign of b indicates the direction of change. Our b of $-568,000$ tells us
that for every successive year the Y_c, the number of horses and mules,
decreases by 568,000. Thus, we can visualize the second Y_c on the vertical

line rising above the X value of 1 as being 568,000 below the starting point we have located. When these two points have been located, the line is determined since one can project it through and beyond these points. The student should practice visualizing geometrically equations written in this form so that the constants a and b will become meaningful to him. We may note that the average annual decrease in horses and mules on farms determined from the least squares trend line, 568,000, is reasonably close to our earlier estimates of 571,000 and 503,000.

It is possible to get a "computed Y" value for each year by reading it from Figure 23 or, computed more accurately, by substituting in equation (4). These Y_c values are often called trend values and are sometimes referred to as "normal" values, since they may with the reservation mentioned on pages 170-171 be interpreted to mean the number of horses and mules on farms in those years if cyclical changes and irregular and accidental happenings had not been present. It is not customary to compute the individual Y_c values since the equation is a more concise way of giving this information.

Thus, we have analyzed the time series of number of horses and mules on farms in the United States from 1918 to 1950 to the extent of describing the secular trend manifested in that period (on the assumption that the trend is linear in form). The results of the analysis can be stated in equation (4), but usually they are also presented graphically as in Figure 23. Again we emphasize that the long-time trend line can be fitted by any one of the several methods explained, or a moving average can be used to show the secular movement with most of the fluctuations removed. However, all of the methods except the fitting of the line by the method of least squares require graphic or detailed tabular presentation, whereas the least squares line can be expressed in one simple, efficient, mathematical sentence.

METHODS OF DESCRIPTION OF SECULAR TREND: NONLINEAR FORM

Nonlinear trends. It may be that the "average annual change" measured by b is not a valid concept for a particular time series. This will be true when the average annual change is considerably greater in one part of the period covered than in another. If so, the fitting of a straight line to describe the trend is not advised, for the straight line describes the progress of a movement where the amount of change per period is constant over the whole range of time observed. Instead of using a straight line in such a case, it is better to try to find a curve which will fit the observed points better. Again such a curve may be drawn in free hand from inspection, or a moving average may be used to approximate the shape of a

curve with most of the irregular fluctuations removed. Finally, an equation may be derived for a geometric figure which is not a straight line. A number of forms of curves which can be used to describe nonlinear trends are shown in Chapter 22, page 447. Looking at the graphic representation of the number of tractors on farms in Figure 20, it can be seen that some sort of curve would describe this trend better than a straight line would.

Growth curves. One group of trend curves of interest to sociologists are those used when the time series embraces a cumulative quantity. Such curves are known as "growth curves." It is obvious that the straight line is poorly adapted to describing an entire history of growth, for neither organisms nor demographic or cultural phenomena increase by constant increments steadily and unchangingly. One of the more frequently used growth curves is the *logistic* curve, sometimes known as the Pearl-Reed curve.[8] Another frequently used form of growth curve is the Gompertz curve.[9]

The problem of curve fitting. For any set of observations, each of which has reference to two continuous variables (as, for instance, to time and to amount of some characteristic), each observation can be plotted as a point on a coordinate system with the axes representing the two variables. When this is done we have a graphic representation of the two sets of observations such as that shown in Figure 20. The procedures of "fitting a curve" to such observations are two: (1) deciding from inspection of table or chart or from other considerations what general form the curve shall have—straight line, logistic, or other; (2) determining from the observed data the value of the constants which when substituted in that general form will define the curve of that form which fits the observed points "best" (as defined by some criterion).

Usually the fitting of curves other than straight lines is treated only in inductive statistics because their use is often accompanied by some sort of generalization of the results beyond the range of observation, for prediction or other purposes. For example, we could extend the least squares straight line computed in this chapter and predict the future numbers of horses and mules on farms in the United States. The fact that this line "predicts" there will be *no* horses and mules on farms in the United States in 1963 makes us somewhat skeptical of its "predictions." The extension of a fitted curve beyond the range of observations is an example of induction and caution must be exercised in such a process. However, there is no reason why the procedures of curve fitting cannot also be used for a purely descriptive and summarizing function.

---

[8] For methods of fitting the logistic curve see *ibid.*, pp. 452–461.

[9] For methods of fitting the Gompertz curve see Charles C. Peters and Walter R. Van Voorhis, *Statistical Procedures and their Mathematical Bases* (New York: McGraw-Hill, 1940), pp. 435–441.

SUGGESTED READINGS

Brumbaugh, Martin A., and Kellogg, Lester S., *Business Statistics* (Chicago: Richard D. Irwin, 1941), Chap. XXI.

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chaps. 14, 15.

Davis, Harold T., *The Analysis of Economic Time Series* (Bloomington, Ind.: Principia Press, 1941).

Deming, W. Edwards, *Statistical Adjustment of Data* (New York: John Wiley and Sons, 1943).

PART III

Inductive Statistics
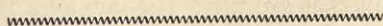
Introduction to Inductive Statistics

Necessity for differentiation of bodies of method on the basis of their function. Perhaps the most important idea to be grasped at this stage of the study of statistics is the differentiation between descriptive and inductive statistics, with the realization that the function to be served in a particular investigation determines the type of statistical methods to be used. In Part III we shall be greatly concerned with methods of estimating for a large group the values of the summarizing measures we have already learned to compute for distributions of characteristics among an observed group of units; with methods of estimating the precision of these estimates; and with methods involving the concept of probability for testing hypotheses about the universe. But of equal importance to the mastery of these methods is the understanding of when to use them and when not to use them. This requires also a knowledge of what assumptions and approximations are involved in generalizing in any particular research problem, and a realization of the necessity of an explicit statement of these assumptions and approximations in the reporting of the results and conclusions. Sociologists on the whole have been lax in these two requirements. The application of these methods in sociological research must become more careful and rigorous if sociology is to earn the right to be called scientific in its use of statistics.[1]

Sociology lags in application of inductive statistical methods. The logic, the theory, and the methods of statistical induction have become rather highly developed through the work of many scholars in pure and applied statistics. The application to sociological research of the theory and methods developed in other fields has not proceeded nearly so far.

---

[1] Samuel A. Stouffer has suggested that we "declare a moratorium on the use of the word 'science' as applied to studies of social phenomena," leave the word to the physical scientists, and be "free—free to borrow as we like from their logics and techniques, and to add logic and techniques of our own, without having to waste time in the scholastic futilities of discussing about a piece of research, 'Now is this science or isn't it?'" "Sociology and Sampling," in L. L. Bernard (ed.), *The Fields and Methods of Sociology* (New York: Long and Smith, 1934), pp. 486–487.

Literature on the principles and practice of applying these methods to sociological research is scant. Excellent examples of application which can be recommended as models are rare. The lack of a rigidly controlled experimental situation in sociological research introduces questions as to the applicability of methods which have not been satisfactorily answered and on which there is no unanimity of opinion among sociological statisticians. Conventions of usage have not had time to become established in the application of the newer methods. One point on which sociological statisticians do agree is that satisfactory answers have not yet been discovered to many of the questions arising.

Requirements in training for application of methods of inductive statistics. This does not mean that one is justified in despair over the situation. It is the belief of the writers that tremendous progress in the development of the application of the methods of general statistics can be made if those sociologists now being trained to do quantitative research have the field clarified for them in the following aspects: (1) the logic, the general methods, and the specific procedures of statistical induction; (2) the conditions and requirements for the "ideal case" for induction; (3) the nature and limitations of the approximation to the "ideal case" in the actual sociological research situation.

If the people who are beginning quantitative research in sociology understand and face these difficult matters, then we can expect that modifications in the design of sociological research can gradually be developed to make the situations more closely approximate the "ideal case"; perhaps that accumulated efforts will empirically establish bases of justification for certain unavoidable approximations; and finally, that valid conventions regarding some of the most difficult points in application can be established as standards to guide research efforts.

Organization of Part III. In attempting to supply the clarification that is prerequisite to further progress, we shall deal in this part, first, with the general nature of induction (Chapter 13); then, with certain fundamental statistical concepts basic to statistical induction (Chapter 14); then, with the actual procedures for estimating universe parameters to describe single quantitative and nonquantitative distributions (Chapters 15 and 16); next, with the practical difficulties met in approximating the ideal conditions in sociological research (Chapters 17 and 18); and finally, with the application of sampling theory in testing hypotheses relating to two distributions (Chapter 19).

Even if these topics were presented as adequately as is possible in the light of the present stage of knowledge, such a presentation would not afford a complete guide for the person beginning quantitative research in sociology. But it is hoped that the presentation here will point up the unsettled questions, will show what approximations are being used and

enable the researcher to recognize them and the limitations they involve, and, finally, will suggest what are the most troublesome points on which immediate efforts should be focused to invent solutions. There is no lack of statistical theory, for mathematical statistics has provided more theory than has been applied in sociology and perhaps more than is applicable. The phase of quantitative research in sociology that is begging for constructive inventive effort is that of correct application.

Induction and Estimation

INVALID inductions of statisticians have been subjected to such stinging and scornful criticisms that many people have become afraid of this logical process for arriving at knowledge and even more afraid of the word "generalizing." Let us examine briefly the meaning of induction, its uses, the specialized methods for making inductions about quantitative phenomena, and their place in statistical research.

Induction is the process of inferring information about a large class of phenomena from the observation of one or more items of the class.[1] It is sometimes loosely defined as "going from the particular to the general," in contradistinction to deduction, "going from the general to the particular."

Nonstatistical inductions. A great part of the everyday knowledge of the world about him which the ordinary individual possesses is the result of induction, either on the part of the individual or on the part of some of his cultural ancestors. Most of the inductions supplying such knowledge—almost all of them except those based upon scientific research —have been made with little or no awareness of the operation of the logical process called induction. Such inductions are made with little attention to the formal requirements for insuring validity of the process. All of us in our every day lives continually make tentative inductions based on one or a very few observations.

For instance, a person having moved to a new city may wonder at what hour a public school one-half block away closes. If on one single afternoon at a few minutes after 3:30 he sees 20 children go past his window in the direction away from the school, he may infer that 3:30 is the hour when the children are dismissed every school day afternoon. The induction may or may not be correct—the hour of dismissal might have been at 3:00 and these 20 children might have been kept in after school as punishment or detained for other reasons. He may make the induction

---

[1] The large class is usually called the "population" or the "universe." For sociological situations the latter term seems preferable since "population" is used in other senses.

only tentatively at first, but if for 20 or 30 week days he observes children passing his window at the same time, his confidence as to the validity of his induction increases.

The new resident might not have waited to learn from observation the hour of closing of the school but might have phoned the principal the first morning and asked him. If on the basis of the principal's statement of the regular hour of dismissal, the new resident anticipated seeing children pass his window at a certain time that afternoon, he would have been employing the process of deduction. Or in the first case after he generalized from his observation as to the regular hour of dismissal, he might have deductively predicted the time when children would pass his window on any particular afternoon.

The two processes are usually woven into an intricate pattern in any complex mental functioning, but the pattern can be analyzed into its identifiable parts.[2] It is valuable to examine these isolated parts, even though, as in any analysis of processes, the atomization of segments is an abstraction. The purpose of the analysis here is simply to point out that the practice of induction is common among nonacademicians who may not be aware of the process.

Induction is also common among academicians engaged in nonquantitative research. Samuel A. Stouffer cites as an example of induction in historical research the fact that the South was suffering from malnutrition at the close of the Civil War, an induction which is very generally accepted as valid, although it is based on only fragmentary samples.[3]

The place of induction in scientific research. As we approach the field of scientific research, we do not find unanimity as to the sequence of processes involved in the acquisition of new scientific knowledge, or as to the primacy of induction or deduction. A group of writers on the logic of scientific research hold that all new knowledge is gained by induction. This is the point of view traditionally held by scientific workers, particularly by those engaged in laboratory experimentation. A statement by R. A. Fisher illustrates the views of those who stress the primacy of induction.

Inductive inference is the only process known to us by which essentially new knowledge comes into the world. To make clear the authentic conditions of its validity is the kind of contribution to the intellectual development of mankind which we should expect experimental science would ultimately supply.[4]

---

[2] Although in the single illustration above, the process labeled an "induction" actually includes an implied deductive process. See Robert Emmet Chaddock, *Principles and Methods of Statistics* (Boston: Houghton, 1925), pp. 25–26.

[3] "Statistical Induction in Rural Social Research," *Social Forces*, 13 (May 1935), p. 505.

[4] R. A. Fisher, *The Design of Experiments*, 5th ed. (New York: Hafner, 1949), pp. 7–8.

On the other hand, there is a group including people in such varied fields as Albert Einstein and J. F. Brown who hold a different point of view about the sequence and importance of processes in scientific research. The following excerpts from Brown's writing illustrate this point of view.

More recently methodological studies have shown us that important scientific discoveries are never made by pure induction. The true method of science is rather the hypothetico-deductive method. . . .

In the hypothetico-deductive method, one must, to be sure, start with experience; but before one can measure, one must have a "hunch" as to what the possible laws of experience may be. . . . The steps in the hypothetico-deductive method become: (1) one gets a hunch about nature; (2) this hunch is formulated into a working hypothesis (*i.e.*, law); (3) the law is verified in an experiment; (4) it is then possible to repeat in a variety of situations the experiment which uncovered the law (*i.e.*, one can make measurements). . . .

There are two ideas of what constitutes the scientific method: induction and the hypothetico-deductive method. For methodological reasons social psychology should use the hypothetico-deductive method. . . .[5]

A similar theme runs throughout Einstein's account of the development of physics, as can be noted from the following selected sentences.

Physical concepts are free creations of the human mind, and are not, however it may seem, uniquely determined by the external world. . . .

Fundamental ideas play the most essential role in forming a physical theory. Books on physics are full of complicated mathematical formulae. But thought and ideas, not formulae, are the beginning of every theory. The ideas must later take the mathematical form of a quantitative theory, to make possible the comparison with experiment. . . .[6]

A further discussion of these ideas may be found in F. S. C. Northrop's *The Logic of the Sciences and the Humanities.*[7]

It is possible that these two points of view may not be so contradictory as they seem, that their difference is in the matter of emphasis on different aspects of research, neither of which precludes the utility of the other. Brown's hypothesis, it seems, is formed by the combination of deductions and tentative inductions, as are most of the formulations ascribed to insight, intuition, and other semi-mystic sources. A hypothesis is a guess, or a statement of a possibility, we grant, but actually it must be based on some experience with the phenomena to which it relates— experience here meaning either direct observation or information secured

[5] J. F. Brown, *Psychology and the Social Order* (New York: McGraw-Hill, 1936), pp. 31–32, 41–42.

[6] Albert Einstein and Leopold Infeld, *The Evolution of Physics* (New York: Simon and Schuster, 1938), pp. 33, 291.

[7] F. S. C. Northrop, *The Logic of the Sciences and the Humanities* (New York: Macmillan, 1947), Chaps. 1 and 2.

through hearing or reading of the observations of others. There may be other processes involved, such as analogies and deductions, but unless we subscribe to the belief that some extrahuman power places new ideas in the mind of man, we see that the formulating of a general hypothesis of principle or law from which specific hypotheses are to be deduced and tested must involve induction. Brown's emphasis is on this induction which is less formally made, at least at the stage of development of research in his field of subject matter, while Fisher's emphasis is on the experimental research and the making of inductions from observations which test the specific hypotheses. Incidentally, Brown's point is well made that the development of measurement techniques must precede this verifying experimental phase of scientific research.

The writers hold with Brown in using the term "scientific research" to cover all these phases in the sequence of processes involved in gaining knowledge but hold with Fisher in being primarily interested here in the last phase, since it is in this phase that statistical methods are more frequently appropriate. If a specific hypothesis deals with quantitative aspects of the phenomena, the methods of general or inductive statistics are indicated for the testing of the hypothesis in the light of observed results. In this connection, some of the anti-statistics proponents have charged that the statistical aspects of research are boring, tedious, that they only prove or verify propositions already known intuitively, and, therefore, that they should not be engaged in by those with higher creative potentialities but should be left to clerks. There may be a part truth in this allegation, but it is often also true that the statistical analysis of observational data studied to test one hypothesis reveals other information which had not been suspected or anticipated and which forms the basis of new hypotheses. That is, not only intuition but also careful study of analyzed quantitative data may give the next research lead in pursuing a problem. We must keep in mind this function of statistics as well as the function of testing tentatively held hypotheses. Both are aspects or phases of scientific research, as are also certain nonstatistical processes designated by Brown. We do not hold with the statisticians who have taken a narrow view of scientific research and who have labeled as nonscientific any method which is nonstatistical. We admit that quantitative methods, since they afford measures of their own precision, can be applied more easily and with a measurable confidence to obtain verifiable information about those phenomena to which they are applicable. But we grant that the more difficult and less well charted methods for dealing with other sorts of phenomena may also be made rigorous, so that the information obtained is verifiable and therefore "scientific."

Statistical inductions. Under certain conditions, when an induction is to be made involving quantitative phenomena, either to formulate or

to test a hypothesis, statistical methods are available and if they are used, the induction is called a statistical induction. The making of a statistical induction usually includes two steps—the making of estimates of values of certain summarizing measures for the universe from which the observed sample was drawn and the making of estimates of the range of error of the estimates of the universe measures. This latter class of estimates includes measures usually called measures of reliability, unreliability, error, precision, or confidence. It is in the second step that statistical inductions have an advantage over other inductions—they not only afford estimates of universe values, but they afford estimates of the accuracy and precision of the particular induction made in terms of probability.

This brings us to the subject of probability and a host of associated notions which differ greatly from the older absolute conceptions of existence and causation of phenomena. While the origin of the use of probability lies in the field of analysis of observational errors, and hence suggests that only those workers in the "not-exact" sciences have to use probability because of lack of precision in measurement, the suggestion is not true, for the newer concepts of quantum physics and relativity can be formulated only in terms of probability.[8] We shall be forced to postpone any detailed discussion of probability and its application in statistical induction until the actual procedures involving its use have been explained.

Reliability and validity of statistical inductions. In nonstatistical inductions one places more confidence in the accuracy of generalizations made on the basis of a large number of observations than in those made on the basis of only a few. But there is no way to judge how much more confidence is justified by, say, doubling the number of observations, or how much less by halving them, or of estimating what range of error might be expected. In statistical inductions, however, we have methods of expressing degree of unreliability of estimates in terms of probability. Therefore, the interpretation of the results of a quantitative research project is less subjective, and, consequently, the statistician can state more precisely both what his induction is and how trustworthy it is. Other things being equal, the margin of error will be smaller the greater the number of cases. But the interpreter of statistical inductions need not feel subject to criticism on the score of too few cases if his margin of error, which has taken the number of cases into account, is small enough for the purposes of his investigation, *and if* the conditions for using probability

---

[8] "The laws of quantum physics are of a statistical character. . . . . Quantum physics formulates laws governing crowds and not individuals. Not properties but probabilities are described; not laws disclosing the future of systems are formulated, but laws governing the changes in time of the probabilities and relating to great congregations of individuals." Einstein and Infeld, *op. cit.*, pp. 299, 313.

theory have obtained. This brings us to the most crucial problem of the application of inductive statistics to sociology—when do practical research situations meet the requirements for the use of probability theory?

It is obvious that even in nonstatistical induction the process of generalizing to a universe from a sample of it will be valid only if that sample is roughly a miniature of the parent universe. We speak of a sample's duplicating the features of its parent universe on a smaller scale as "representativeness" of the sample. Representativeness is a quantitative characteristic possessed in varying degrees by samples, not an all-or-none attribute. It is also a composite characteristic, for a sample may be representative of its parent universe in the distribution of one characteristic but not in the distribution of another characteristic. There are various ways of securing representative samples, but the method preferred for statistical induction is that of random sampling, simple or stratified, which will be defined and discussed in Chapters 18 and 19.

Situations requiring descriptive statistics. Let us consider now the situations in sociology requiring the different methods of statistics. It seems that there is a rough parallel between the differentiation between social surveys and social research made by Pauline V. Young [9] and the differentiation between descriptive and inductive statistics (in its second use described below). If a survey has as its purpose the description of a set of conditions for a certain area at a certain time, usually with the implied purpose of measuring them in order to plan some program of reform to alter them, then descriptive statistics may be adequate for the analysis of such quantitative data as are collected in this type of project.

Practical sampling situations requiring inductive statistics. If, on the other hand, either of the following situations is the case, inductive statistics is indicated. The first situation is the same as that described above for the survey, except that the area is too great for every unit to be counted or measured, and only a sample of units is observed. Here one uses inductive statistics to form estimates of various summarizing measures for the whole area (universe) with accompanying measures of unreliability of the estimates. This function of inductive statistics is simply a substitute for descriptive statistics when all units cannot be surveyed.

Hypothetical sampling situations requiring inductive statistics. The other use of inductive statistics is not so easy to state or explain. In this second situation, from observations made either from a sample or from complete survey of some limited universe, we generalize to a hypothetical universe which is difficult to define. It is the universe of all possible samples (which may be limited universes) which could have been produced under similar conditions of time, place, culture, and other relevant fac-

---

[9] *Scientific Social Surveys and Research: An Introduction to the Background, Content, Methods, and Analysis of Social Studies,* 2d ed. (New York: Prentice-Hall, 1950), p. 62.

tors. Generalizing to such a hypothetical universe provides the sociologist with hypothetical "universals," which are not "universal" in an absolutist sense, however. They differ from the universals similarly arrived at by chemists, physicists, and others in a greater degree of complexity in the specification of the "similar conditions." A chemist can state as a "universal" that under specified conditions of temperature, humidity, pressure, etc. such and such a chemical reaction will take place in such and such a fashion when certain elements are brought together. His statement will be true "universally," which means in ordinary thought, regardless of time, location, or the culture or civilization of that location. Since, however, the phenomena which interest sociologists are affected by time, location, the nature and stage of a culture, any sample taken at one time in one area and observed as to phenomena associated with culturally conditioned human beings cannot be considered as representative of all time, all locations, all cultures. Without considering further the existence or meaning of universals in physical sciences, we can say that no one knows whether or not real—in contradistinction to hypothetical—sociological universals exist. Certain societal processes such as competition have been observed in many different time-place-culture combinations, and nonstatistical generalizations have been made as to their universality. But the practical limitations of securing data representative for all societies at all stages in all times is beyond the scope of any immediate possibility in the present stage of sociological research. For the present, at least, we shall have to content ourselves with "universals" whose universality is greatly diminished. For certain phenomena on which we have observations extending over a period of years, we can broaden the time span for which a generalization holds. But for most research projects which cut a cross section in time our generalizing is to the hypothetical universe of such universes as could have been produced within the stated conditions. It may help to imagine an expanded or prolonged present, where the dynamic forces continue to operate as of the specified time, producing phenomena which show only "chance" variation.[10]

To an absolute determinist this conception has no validity. To him the observed results are the only ones that could have been caused and therefore could have been observed. To those who are attempting to integrate into the theory of causation of sociological phenomena the newer ideas of indeterminacy expressed in probabilities, the procedure of making a statistical induction in such a research situation seems to be a promising method. The matter is somewhat controversial, partly because the idea as well as the utility of the concept of such a hypothetical universe is relatively new and not too well defined.

[10] For further discussion of "chance" as a name for factors producing the variation observed between random samples drawn from the same universe, see p. 223 below.

Barring for the present the possibility of establishing and describing truly universal universals in sociology (they are at least barred from quantitative research, although the organismic, philosophy of history, or other types of generalizing may arrive at them by means of nonstatistical inductions), it seems that we have to accept this concept of generalization to a hypothetical universe of possibilities if we wish to proceed beyond the stage of mere description of a unique set of data toward the unraveling of the order or regularity underlying varied manifestations observed. The path is not very clear. We shall explore it further after the procedures for making statistical inductions have been explained. Perhaps more tools are available than we can use meaningfully. Yet careful and critical use of the ones which appear to be appropriate may lead to a clearer understanding of their full meaning. The eventual test of the usefulness of these statistical tools will, of course, be the fruitfulness of the results they yield.

Processes involved in statistical induction. Let us examine more specifically the processes involved in making the most commonly used statistical inductions and the assumptions on which they are based. The steps can be listed as follows:

1. The definition or postulation of a universe, existent or hypothetical, about which we wish to gain knowledge.

2. The selection of a sample from the universe in such a way that the assumption is justified, or approximately justified, that only chance factors make this sample different from any other sample of equal size similarly drawn from the same universe.

In the case of sampling from a hypothetical universe in sociological research, it is not completely clear what this assumption means. For instance, if we start with a complete enumeration of a group of units as the sample, the postulated universe of possibilities from which this actual group of units may be considered a sample is a mental or logical construct which may or may not be worth generalizing to.

3. Enumeration or measurement of the units of the sample with regard to one or more characteristics.

4. Analysis of the quantitative data recorded about the units of the sample to yield estimates of various features of the distribution or distributions of the characteristics in the universe.

5. Computation of measures to be used to evaluate the accuracy or reliability of the above estimates.

6. Testing of hypotheses about the universe by the use of the estimates derived in 4 and the measures of reliability computed in 5. Such tests are often called tests of significance. They make possible the testing of a hypothesis about the universe from the measures derived from observations on the sample. Such tests are limited to disclosing the probability that the results obtained would be observed in a random sample drawn from a universe which has certain speci-

fied values of its summarizing measures. If the probability is small, we deem the hypothesis about the universe untenable and reject it. If the probability is moderate or large, we can merely say that our results are in accord with the hypothesis—not that they prove this particular hypothesis, for they might be in accord with other hypotheses as well. A special case of these tests of hypotheses is the establishing of confidence limits around an estimate, where the hypotheses are implicit.

Limitations and utility of the methods of inductive statistics. Thus we see that the methods of inductive statistics are admittedly limited in their revealing of information about the universe and proving its validity. They enable us to make the best guesses about the universe from the information at hand, to estimate the reliability of the guesses, to show that certain guesses are highly improbable, or that others are not inconsistent with the observed facts. But statistical methods offer no final or positive proof of any hypothesis. Nevertheless, careless and unjustified methods of dealing with quantitative characteristics showing variability have prevailed for so long that the mere negative function of statistics in demonstrating the caution which needs to be exercised should be valuable to an embryonic science groping to advance beyond the purely speculative stage. And when to this function is added the positive function of increasing the fund of information which, because of the precision of the statistical tools by which it was gained, is verifiable by others and is established as securely as is justified by the observational data on which it is based, the utility of these methods appears evident.

SUGGESTED READINGS

Cohen, Morris R., and Nagel, Ernest, *An Introduction to Logic and Scientific Method* (New York: Harcourt, 1934).

Churchman, C. West, *Theory of Experimental Inference* (New York: Macmillan, 1948), Chaps. 1, 2.

Fisher, R. A., *The Design of Experiments*, 6th ed. (London: Oliver and Boyd, 1951), Chap. 1.

Larrabee, Harold A., *Reliable Knowledge* (Boston: Houghton, 1945), Chaps. 11, 12, and 13.

Northrop, F. S. C., *The Logic of the Sciences and the Humanities* (New York: Macmillan, 1947).

Stouffer, Samuel A., "Some Observations on Study Design," *The American Journal of Sociology*, LV (January 1950), pp. 355–361.

‹‹
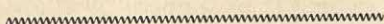
The Normal Curve

The importance of the normal distribution. In any exposition of the methods of statistical induction one will meet so frequently phrases such as "a normal distribution" or "normally distributed" that it seems necessary to digress from the presentation of the methods of making inductions and testing hypotheses to examine what is the most important form of distribution in statistics, the normal distribution. A noted statistician has paid the following tribute to the normal distribution.

Many years ago my inspiring teacher, Henry Lewis Rietz, observed that statistical procedures derived from the normal distribution were born of a higher realm than other procedures. I disbelieved this with a religious fervor. Though, as time has passed, I have espoused curvilinear regression and skew distributions with gusto, I have found myself frequently slipping, for the data would not support me, and linear relationships and nearly normal distributions have in my experience as a psychologist cropped up with a frequency which has chided and mocked me. I still reserve judgment as to the place of birth of the normal distribution, but that its sphere of usefulness is extended in connection with biological and psychological phenomena I no longer have the slightest doubt.[1]

The history of the successive discoveries about the normal distribution is a fascinating account of contributions by gamblers, mathematicians, and others, excellently summarized by Helen Walker in *Studies in the History of Statistical Method*.[2] There are also several approaches to the mathematical derivation of the normal curve, both theoretical and empirical, which can be found in Thornton C. Fry's *Probability and Its Engineering Uses* [3] and in texts of mathematical statistics.

One of the more interesting origins of the normal curve is in the dis-

---

[1] From Truman Lee Kelley, *The Kelley Statistical Tables*, rev. ed., (New York: Macmillan, 1948), Preface. By permission of the Macmillan Company, publishers.
[2] Helen Walker, *Studies in the History of Statistical Method with Special Reference to Certain Educational Problems* (Baltimore: William and Wilkins, 1929).
[3] Thornton C. Fry, *Probability and Its Engineering Uses* (New York: Van Nostrand, 1928).

tribution of the deviations from an expected value. For example, let us take a universe of measurements and consider the deviations from the average or expected value. Let us suppose "that any deviation is the result of the operation of an indefinitely large number of small causes, each producing a small perturbation. Let us assume that the small perturbations are all equal, and that positive and negative perturbations are equally likely." The resulting distribution is normal.[4]

Our treatment of the normal curve in this chapter will be limited to a description of the aspect of form of the normal distribution, some useful relations between its abscissas, areas, and ordinates, certain tables based upon these relations, and methods of comparing the form of an observed distribution with the form of the normal distribution as a "standard." In the chapters following this one, we shall consider applications which statisticians make of the form and the relations of abscissas, areas, and ordinates of the normal distribution.

DESCRIPTION OF THE NORMAL DISTRIBUTION

The form of distributions. We have repeatedly stated that the data amenable to statistical analysis are those resulting from enumeration or measurement of a series of units varying from one another with respect to an enumerable or measurable characteristic. Such data record the distribution of the characteristic. Distributions of quantitative characteristics vary in central tendency, in dispersion, and in form. Since summarizing measures of central tendency and dispersion are usually given in terms of the units of measurement of the characteristic, they may be greatly different for different characteristics. In regard to the aspect of form, however, for which we have not yet presented precise summarizing measures, a great many observed distributions of characteristics are quite similar. Even more similar in form are many of the "sampling distributions" with which we shall deal in the next few chapters. The form of distribution which is approximated so frequently that it comprises the "standard" of form of distributions is the normal distribution.

Frequency curves. The curve of a quantitative distribution is the limiting form of either the histogram or the coordinate chart representing the frequency distribution resulting from grouping, when the size of the class interval becomes infinitesimally small and the numbers of cases and of class intervals become infinitely great. The curves approached by histograms and coordinate charts constructed from sample data are used to picture the estimated distribution of a characteristic in an unlimited

---

[4] See G. Udney Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), p. 187.

universe.[5] They are the graphic representations of generalizations to a universe with respect to the distribution of a characteristic. Often an equation can be determined which describes algebraically the distribution in the universe. We frequently describe the form of an observed distribution by reference to the curve approached by its histogram. Similarly, we
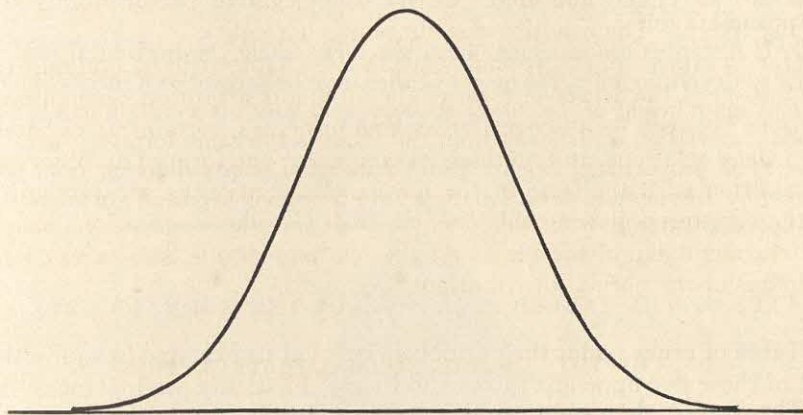
Figure 24.   The Normal Curve.

shall describe the form of a normal distribution by describing a normal curve.

Form of the normal curve.   The adjective "normal" refers only to the aspect of form of a distribution. Remembering that the form of a distribution is not dependent upon the number of cases or the units of measurement, let us examine the simpler characteristics of a normal curve without regard to any units of measurement. The normal curve shown as Figure 24 has the following features.

1. It is an *I*-type curve with its greatest frequency in the center and its tails approaching the horizontal axis asymptotically.

2. It is symmetrical about the mean. There are as many cases above as below the mean, which makes its median coincide with its mean; the greatest frequency is at the mean, which makes its mode coincide with its mean. Its skewness is therefore zero.

3. It has a specified degree of kurtosis or peakedness, which we shall learn to measure shortly.

Relations of the normal curve.   In the normal curve there are two important relations, one between area and abscissa, the other between ordinate and abscissa. The relations are as follows:

---

[5] The concept of a "universe" and the process of generalizing to a universe is discussed in more detail in the next chapter.

1. If distances are measured from the mean along the horizontal axis in standard deviation units, the proportion of the area lying under the curve between the mean and a point which is a certain number of standard deviation units away from the mean is the same for every normal curve. The proportions of area found between the mean and successive points differing from one another by .01 standard deviation units have been computed and are tabulated in Appendix Table A.

2. If distances are measured from the mean along the horizontal axis in standard deviation units, the height of the curve (expressed as a proportion of the maximum height of the curve) above a point which is a certain number of standard deviation units away from the mean is the same for every normal curve. The proportional heights above successive points differing from one another by .01 standard deviation units have been computed and are tabulated in Appendix Table B.

USES OF THE TABLED RELATIONS OF THE NORMAL CURVE

Table of areas under the normal curve. Let us examine carefully the first of these two appendix tables and its use. First, any original measures of a characteristic must be expressed as a deviation from the mean of the distribution and then transformed into standard deviation units before the tables are used. A measure expressed as a deviation in standard deviation units is simply the difference between the actual measure and the mean divided by the standard deviation, thus,

$$\text{Deviation of a measure in standard deviation units} = \frac{X - \bar{X}}{s} \text{ or } \frac{x}{s} \qquad (1)\,[6]$$

The normal curve theoretically has an unlimited range, since its tails approach but never quite reach the horizontal axis. Yet because over 99 percent of the area under the curve lies between ordinates erected at points three standard deviation units above and three standard deviation units below the mean, tables of areas and ordinates of the normal curve usually do not show values for $\frac{x}{s}$ greater than three or five standard deviation units. Since in statistical writing and in research reporting, one frequently finds references to the area (or to the frequency or number of cases) found within plus or minus one standard deviation from the mean, plus or minus two standard deviations from the mean, and plur or minus three standard deviations from the mean, the student should look up these values in Appendix Table A and compare them with the percentages shown in Figure 25 to fix them in mind.

---

[6] Sometimes written $\frac{x}{\sigma}$; we shall distinguish between s and σ in Chapter 16.

Proportion of cases lying between the mean and another value of a measure. It frequently happens that for a particular distribution of normal form one wishes to know the area or proportion of cases included between the mean and some value of a measure either above or below the mean. To do this one first expresses the difference between the values and the mean in standard deviation units as indicated in formula (1). He then turns to Appendix Table A and finds this value to one decimal place in the leftmost column, called the argument, of the table. With the row lo-
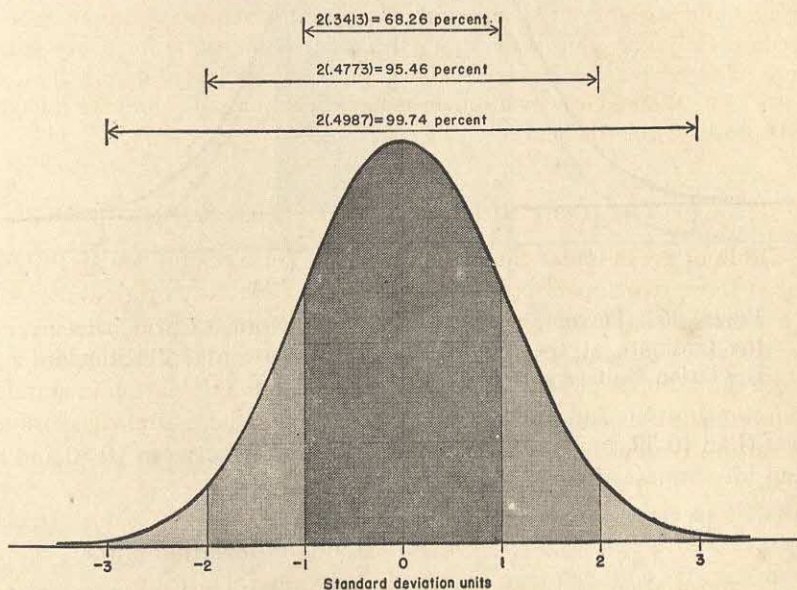
Figure 25. Percentages of Area Falling within Ordinates at Plus and Minus One, Two, and Three Standard Deviation Units from the Mean in a Normal Distribution. (Source: Appendix Table A.)

cated, he goes to the right until under the column heading designating the second decimal place of the value and reads off the entry in the cell so located. This entry is the proportion of the number of cases lying within the range bounded by the mean and the value looked up.

Let us illustrate the procedure of using Table A with data from the actual distribution of number of children borne by the 117 white tenant farm women of Chapter 8. We shall for the moment assume that the distribution of children is normal. We have already found that the mean of the distribution is 6.31 and the standard deviation is 3.45. Now suppose that we know only these figures and do not have the frequency distribution given, and that we wish to know first what proportion and then what number of women we should expect to have more than 6.31 children but
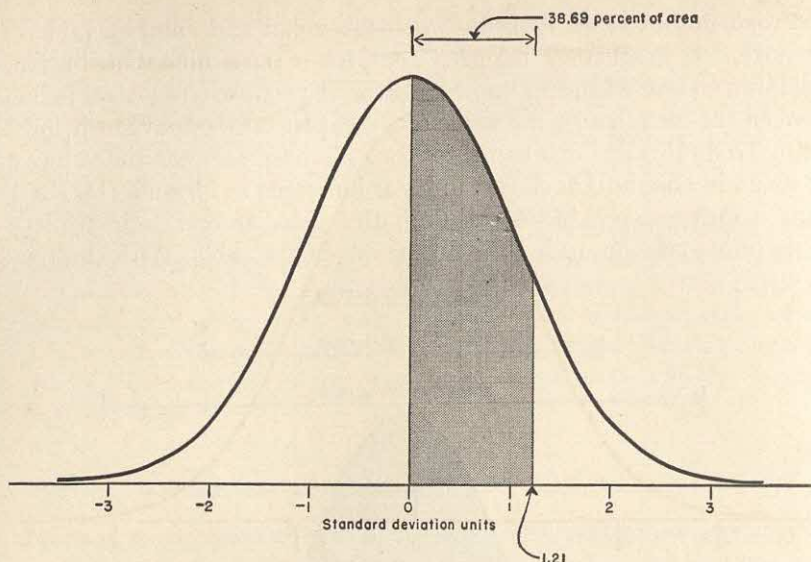
Figure 26.   Percentage of Area Under the Normal Curve between the Ordinate at the Mean and the Ordinate at 1.21 Standard Deviation Units. (Source: Appendix Table A.)

fewer than 10.50. First, we translate the difference between 10.50 and the mean into standard deviation units, thus,

$$\frac{10.50 - 6.31}{3.45} = \frac{4.19}{3.45} = 1.21 \text{ standard deviation units}$$

Next, in the leftmost column of Appendix Table A we find the first two digits of 1.21, that is, 1.2, on the thirteenth row of the table, and we follow this row across the page to the column under the heading .01. Here we find the cell entry, 3869, which is interpreted to mean that .3869 or 38.69 percent of the area of the normal curve lies between an ordinate erected at the mean and an ordinate erected 1.21 standard deviation units above (or below) the mean. Figure 26 shows this graphically. Remembering that the area under a curve is proportional to the frequency or number of cases, we can compute the actual number of women we should expect to have borne between 6.31 and 10.5 children by simply multiplying .3869 times 117, the total number of women. Thus we see that if the number of children borne by these women is a normally distributed characteristic, we should expect 38.69 percent of the 117 women, or 45 of them, to have borne between 6.31 and 10.50 children.

Proportion of cases lying between two values of a measure. Suppose we want to know how many women we should expect to have borne
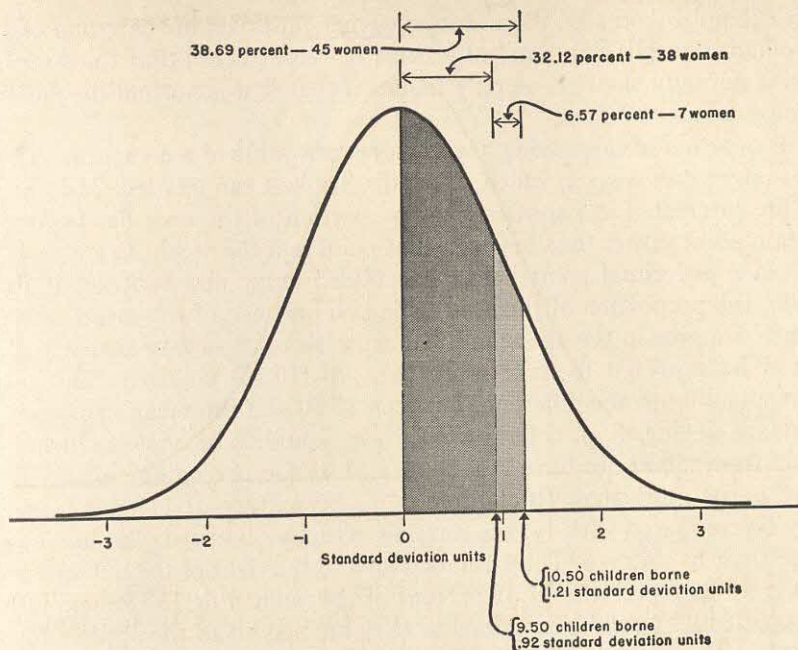
Figure 27. Percentage and Number of Women Expected to have Borne 10 Children. (Source: Table 11; Appendix Table A.)

10 children. It is necessary to remember that in treating a distribution of only integral (whole number) values as a continuous distribution, we must consider the integer "10" as spread over the range from 9.50 to 10.50. There is no way to look up directly in a table the proportion of cases falling between two such values as 9.50 and 10.50. All the table can tell us is the proportion falling between the mean and some other value. Therefore, we find the proportion and then the number expected between the mean and 9.50, and subtract this number from 45, the number already found expected between the mean and 10.50. As above we have

$$\frac{9.50 - 6.31}{3.45} = \frac{3.19}{3.45} = 0.92 \text{ standard deviation units}$$

From the same Appendix Table A we find the entry corresponding to 0.92 is 3212. Again we get 32.12 percent of 117 and find that 38 women is the number expected to have borne between 6.31 and 9.50 children. Then to get the desired information of how many women we should expect to have borne 10 (or between 9.50 and 10.50) children, we subtract 38 from 45 and get 7. Figure 27 shows the areas and corresponding percentages and numbers of women falling within the different ranges referred to in

this example. Actually, there were 6 women observed in the group of 117 who had borne 10 children, but it must be remembered that the distribution is actually skewed; we only assumed that it was normal for illustrative purposes.

Proportion of cases lying beyond a certain value of a measure. There are many other ways in which Appendix Table A can be used. Most often we are interested in knowing what proportion of the area lies beyond a certain point rather than between that point and the mean. In such a case we place a decimal point before the tabled entry and subtract it from .5000, the proportion of the area falling on one side of the mean. For instance, suppose in the above problem we wish to know how many women would be expected to have more than 10 (10.50) children. The entry corresponding to the difference between 10.50 and the mean expressed in standard deviation units has already been found to be 3869. Subtracting .3869 from .5000, we have .1131 or 11.31 as the percentage expected to have borne more than 10 children. This percentage of 117 is 13, which may be compared with 17, the number actually observed who had borne more than 10 children. The unshaded area to the right of the hatched area in Figure 27 represents 11.31 percent of the women or 13 women. Other uses of Table A will be explained as they are needed in the application of sampling theory. Appendix Table A is sometimes called the table of areas under the normal curve, or sometimes the table of half areas.

Table of ordinates of the normal curve. The second table mentioned, Appendix Table B, does not have such extensive use as the first. It is called the table of ordinates of the normal curve, since it gives the proportional heights of vertical lines from the horizontal axis to the curve. Its argument is the same as that of Appendix Table A, standard deviation units by steps of .01, while its cell entries give the height of the ordinate at the point corresponding to the argument, expressed as a proportion of the height of the maximum ordinate, which is at the mean.

Table B is used chiefly in fitting a normal curve to an observed distribution, and this use will be illustrated now, although the reasons for fitting a curve and the meaning of the process of fitting will not be discussed fully until Chapter 15.

Proportion of the area more than a certain number of standard deviations from the mean. Frequently we will need to know the proportion of the area that lies more than a certain number of standard deviations from the mean in either direction. In the preceding paragraph we learned how to find the proportion of cases that lies more than a certain number of standard deviation units from the mean in one direction. We could double this to find the proportion lying more than this number of standard deviation units from the mean in either direction. However, we need this measure so frequently that Appendix Table C has been made to give this
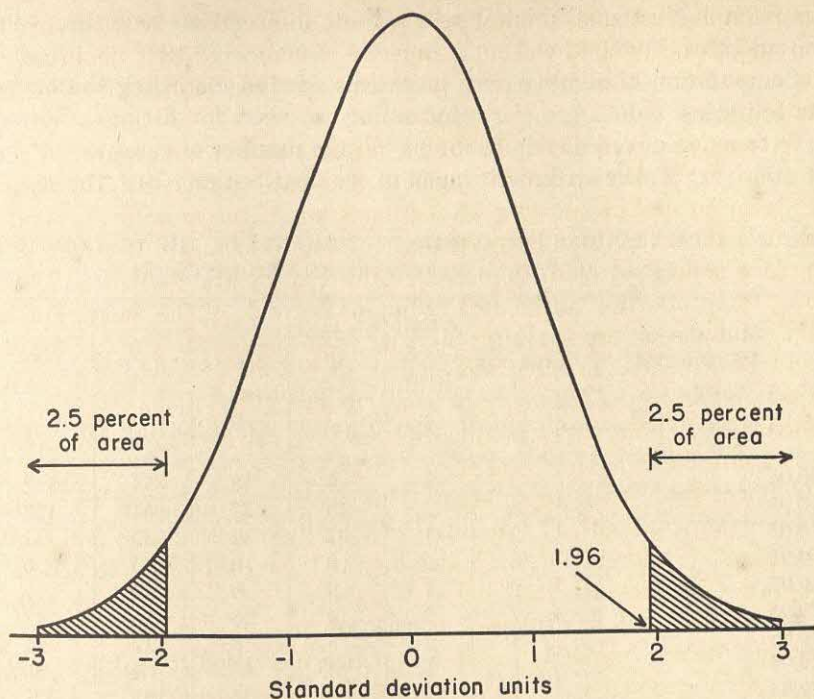
Figure 28.   Percentage of Area under Normal Curve Lying More than 1.96 Standard Deviation Units from the Mean. (Source: Appendix Table C.)

measure directly. Suppose we wish to know the proportion of cases lying more than 1.96 standard deviation units from the mean in either direction. Looking up 1.96 in Appendix Table C, we obtain .0500, which means that .05 or 5 percent of the area under the curve lies more than 1.96 standard deviation units from the mean. This is shown in Figure 28 with 2.5 percent of the area lying more than this distance above the mean and 2.5 percent lying more than this distance below the mean. The ways in which this information is used will be discussed later.

FITTING A NORMAL CURVE TO AN OBSERVED DISTRIBUTION

Without being concerned over why we wish to have a curve fitted to an observed distribution, let us learn the procedures for fitting a normal curve to the data of Table 16 in Chapter 9. We will, however, omit the one extreme case in Table 16 since it has such a major effect on the mean and standard deviation and fit our normal curve to the remaining 99 counties. A research worker must be extremely hesitant to omit a case

because it is "unusual" and should include information regarding such omitted cases.

Computation of summarizing measures needed for fitting the curve. The following values are the information we need for fitting a normal curve to an observed distribution: (1) N, the number of cases in the distribution; (2) $\bar{X}$, the arithmetic mean of the distribution; (3) s, the stand-

Table 26. COMPUTATIONS FOR OBTAINING MOMENTS OF THE DISTRIBUTION OF TABLE 16 AFTER OMITTING OPEN-END INTERVAL [a]

Midvalue of class interval m (1)	Frequency f (2)	Step deviations d' (3)	fd' (4)	$f(d')^2$ (5)	$f(d')^3$ (6)	$f(d')^4$ (7)
−20.05	2	−3	−6	18	−54	162
−10.05	8	−2	−16	32	−64	128
−0.05	32	−1	−32	32	−32	32
9.95	38	0	0	0	0	0
19.95	9	1	9	9	9	9
29.95	7	2	14	28	56	112
39.95	1	3	3	9	27	81
49.95	1	4	4	16	64	256
59.95	1	5	5	25	125	625
Sums	99		−19	169	131	1,405

[a] See p. 130 for discussion of this omission.
Source: Table 16.

ard deviation of the distribution, and (4) i, the width of the class interval. We obtain these measures by the methods of grouped data given in Chapters 8 and 9. Table 26 shows the computations for obtaining the necessary sums. We obtain $\bar{X}$ and s as follows:

$$\bar{X} = \bar{X}' + \frac{\Sigma fd'}{N} i = 9.95 - \frac{19}{99} 10 = 8.031$$

$$s = i \sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} = 10 \sqrt{\frac{169}{99} - \left(\frac{-19}{99}\right)^2} = 12.924$$

Determining constants for the normal equation. The problem of fitting a normal curve to an observed distribution requires the evaluation of certain constants of the general equation for the normal curve, which is,

$$Y_c = ke^{\frac{-(X-a)^2}{2b^2}}$$

$$(2)$$

The general equation can be written in a modified form which indicates clearly just how to find the values of the three constants contained in (2) in fitting the curve to an observed distribution, thus,

$$Y_c = \frac{Ni}{s\sqrt{2\pi}} e^{\frac{-(X - \bar{X})^2}{2s^2}} \tag{3}$$

where N = number of observations
i = width of class interval
s = standard deviation of observed distribution
$\bar{X}$ = mean of observed distribution

Substitution of the values of N, i, s, and $\bar{X}$, obtained from grouped data of an observed quantitative distribution will provide an equation of a normal curve which has the same mean and standard deviation as the observed distribution, and which describes the same number of cases, with reference to the same size interval as used in grouping the observed distribution. Substituting the data for our example in equation (3) gives the equation desired,[7]

$$Y_c = \frac{(99)(10)}{12.597(\sqrt{2\pi})} e^{\frac{-(X - 8.031)^2}{2(12.597)^2}} \tag{4}$$

This can be reduced to the following form,

$$Y_c = 31.353\, e^{\frac{-(X - 8.031)^2}{2(12.597)^2}} \tag{5}$$

Graphic presentation of the fitted normal curve. The algebraic part of the fitting of the curve has been completed, but we usually want a graphic presentation of the fitted curve, ordinarily plotted on a chart along with the observed distribution so that the two may be compared. It is possible to obtain values of Y_c corresponding to selected values for X by straightforward logarithmic evaluation of (5). However, the necessity of such long arithmetic computations has been obviated by the construction of Appendix Table B. The computations have already been performed for successive values of

$$\frac{x}{s} = \frac{X - \bar{X}}{s}$$

and the Y_c values are expressed as proportions of Y_o, the value of Y_c at the mean. Before using the table we evaluate equation (3) for $X = \bar{X}$ which gives us a general equation for Y_o.

---

[7] Instead of substituting $s = 12.924$, the value just obtained, we are substituting a "corrected" value of s, 12.597. The procedure for making the correction is explained on p. 217.

$$Y_0 = \frac{Ni}{s\sqrt{2\pi}} \, e^{\frac{-(\bar{X} - \bar{X})^2}{2s^2}}$$

$$Y_0 = \frac{Ni}{s\sqrt{2\pi}} \, e^0$$

Table 27. COMPUTATIONS FOR OBTAINING ORDINATES
OF THE NORMAL CURVE FITTED TO THE DATA
OF TABLE 26

$\dfrac{x}{\sigma}$	X (1) $\times \sigma + \bar{X}$	Proportional height of ordinates (Appendix Table B)	Y_c (3) $\times Y_0$
(1)	(2)	(3)	(4)
−3.0	−29.760	.01111	.348
−2.8	−27.241	.01984	.622
−2.6	−24.721	.03405	1.068
−2.4	−22.202	.05614	1.760
−2.2	−19.682	.08892	2.788
−2.0	−17.163	.13534	4.243
−1.8	−14.644	.19790	6.205
−1.6	−12.124	.27804	8.717
−1.4	−9.605	.37531	11.767
−1.2	−7.085	.48675	15.261
−1.0	−4.566	.60653	19.017
−0.8	−2.047	.72615	22.767
−0.6	.473	.83527	26.188
−0.4	2.992	.92312	28.943
−0.2	5.512	.98020	30.732
0.0	8.031	1.00000	31.353
0.2	10.550	.98020	30.732
0.4	13.070	.92312	28.943
0.6	15.589	.83527	26.188
0.8	18.109	.72615	22.767
1.0	20.628	.60653	19.017
1.2	23.147	.48675	15.261
1.4	25.667	.37531	11.767
1.6	28.186	.27804	8.717
1.8	30.706	.19790	6.205
2.0	33.225	.13534	4.243
2.2	35.744	.08892	2.788
2.4	38.264	.05614	1.760
2.6	40.783	.03405	1.068
2.8	43.303	.01984	.622
3.0	45.822	.01111	.348

Source: Table 26; Appendix Table B.

$$Y_o = .39894 \frac{Ni}{s} \tag{6}$$

Substituting the values of our example in (6), we have

$$Y_o = .39894 \frac{(99)(10)}{12.597} = 31.353$$

Table 27 shows the procedures for getting the corresponding Y_c and X values. The entries in column (1) are arbitrarily selected; in this example we have chosen to plot a point at each .2 of a standard deviation unit over a range extending from a point located at -3 standard deviation units below the mean to a point located $+3$ standard deviation units above the mean. The methods of computing the successive columns are noted at the head of each column. Columns (2) and (4) of this table contain the corresponding values of X and Y_c for 31 points on the normal curve. In Figure 29 we see these points plotted and connected by straight lines as
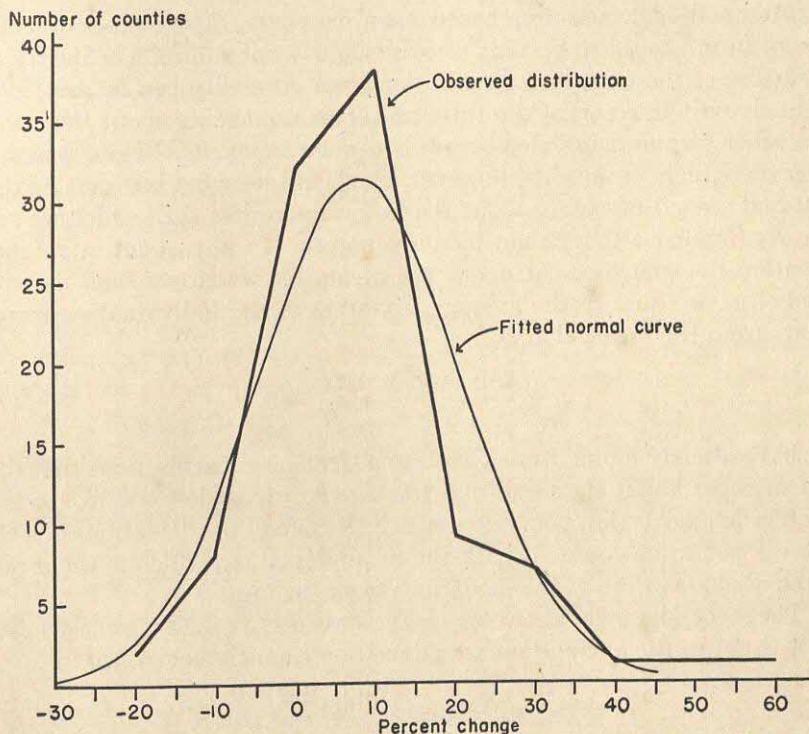
Figure 29. Observed Distribution of 99 North Carolina Counties by 10-unit Internals of Percentage Change in Population, 1940–1950, with Fitted Normal Curve. (Source: Table 27.)

well as connected points representing the observed frequencies in each of the class intervals.

Graphic comparison of the form of the observed distribution with the normal form is possible by inspection of Figure 29. In general the two forms are similar. However, with respect to skewness, we can see that the observed distribution shows a slight skewness to the right (positive skewness) when the absolutely symmetrical normal distribution is used as a standard of comparison. And with respect to kurtosis, we can see that the observed distribution is more peaked (leptokurtic) with longer tails than the normal distribution. Often such a graphic comparison is all that is needed for describing the departure of the form of a particular distribution from normality. By graphic comparison we can usually detect both direction and a rough idea of degree of skewness as well as whether a form is leptokurtic or platykurtic.

SUMMARIZING MEASURES OF FORM

Moments and measures based upon moments. If an exact description of form is needed, graphic comparison will not suffice. The degree of departure of the form of a distribution from normality can be described precisely only in terms of the third and fourth moments about the mean and other summarizing measures based upon moments. Before we consider these higher moments, however, we should see what is meant by the first and second moments about the mean, quantities with which we are already familiar although not by these names. For any quantitative distribution the first moment about the mean, for which we shall use the symbol μ_1, is equal to the average deviation of the individual measurements from the mean, that is,

$$\mu_1 = \frac{\Sigma(X - \bar{X})}{N} = \frac{\Sigma x}{N} = 0 \qquad (7)$$

We have already found in the chapter on frequency distributions that the first moment about the mean of a frequency distribution is zero, for the mean is defined as that point from which the sum of the deviations is zero. (This is not to be confused with the mean deviation, which is the mean of the *absolute values* of the deviations from the mean.)

The second moment about the mean, for which we shall use the symbol μ_2, is equal to the average squared deviation from the mean, that is,

$$\mu_2 = \frac{\Sigma(X - \bar{X})^2}{N} = \frac{\Sigma x^2}{N} \qquad (8)$$

One recognizes immediately that the second moment is simply another name for the quantity we have defined as the variance, or the square of the standard deviation.

Moments higher than the second define quantities with which we have not dealt before but which are useful in describing the form of a distribution. As one might expect, higher moments are defined analogously to the first and second moments. The third moment about the mean, μ_3, is equal to the average of the cubed deviations from the mean, that is,

$$\mu_3 = \frac{\Sigma(X - \bar{X})^3}{N} = \frac{\Sigma x^3}{N} \tag{9}$$

If a distribution is perfectly symmetrical, the sum of the negative x^3's will be exactly equal to the sum of the positive x^3's, and the third moment will hence be equal to zero. This is true in case of the normal distribution, which is absolutely symmetrical. If a distribution is skewed, however, the cubes of the deviations of the measures in the longer tail will not be balanced by the cubes of the smaller deviations in the other direction. Cubing deviations gives even more importance to the greater deviations than squaring them. The sign of the third moment will be the same as the direction of the skew, that is, the direction where there are more extreme values, and the value of the third moment will be greater, the more skewed the distribution is. Therefore, μ_3 may be used as a measure of skewness, although it is an absolute rather than a relative measure since it involves a power of the units of measurement. In order to obtain a summarizing measure of skewness related to μ_3 but which is a ratio rather than a measure expressed in terms of the original units of measurement, the measure, β_1, is defined thus,

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3} \tag{10}$$

Since both the numerator and the denominator of this expression involve the same (sixth) power of the units of measurement, their quotient, β_1, is independent of the units. Sometimes β_1 itself is taken as a measure of skewness, sometimes $\sqrt{\beta_1}$ is taken,[8] and sometimes other expressions involving β_1 are used as coefficients of skewness. Since the third moment of any symmetrical distribution, and hence of a normal distribution is zero, it is obvious that β_1 for a normal distribution is zero. With this as the standard of absolute symmetry, the size of the β_1 of any distribution indicates how far the distribution departs from perfect symmetry of a normal distribution. Figure 30 shows three curves of distributions with different values of $\sqrt{\beta_1}$.

The fourth moment about the mean, μ_4, is equal to the average of the fourth powers of the deviations from the mean, that is,

$$\mu_4 = \frac{\Sigma(X - \bar{X})^4}{N} = \frac{\Sigma x^4}{N} \tag{11}$$

---

[8] The $\sqrt{\beta_1}$ is called γ_1 and is given the sign, plus or minus, of the third moment.
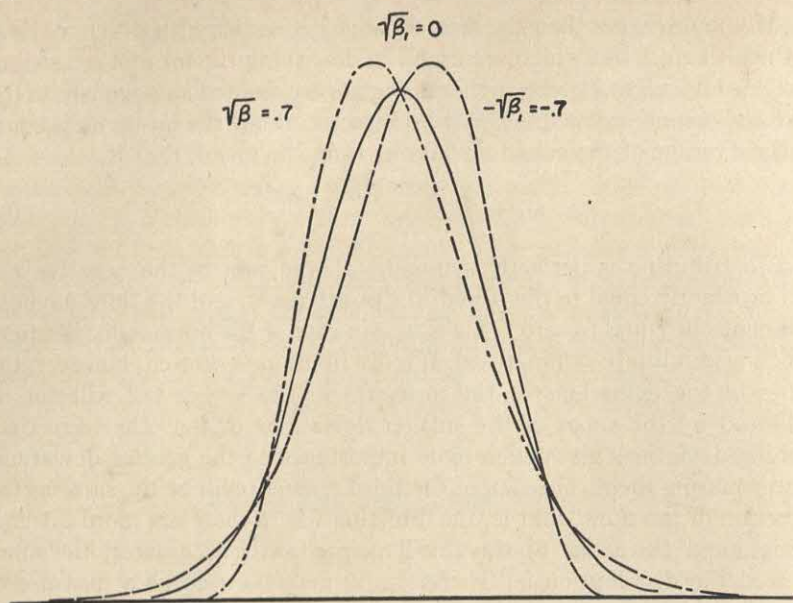
Figure 30.   Curves with Different Skewness.

Since any even power of a deviation is positive, no matter what the sign of the deviation, μ_4 like μ_2 will always be positive. In order to obtain a measure related to μ_4 but independent of the units of measurement, still another summarizing measure, β_2, is defined thus,

$$\beta_2 = \frac{\mu_4}{\mu_2{}^2} \tag{12}$$

Since both the numerator and the denominator of this expression involve the same (fourth) power of the units of measurement, their quotient, β_2, is independent of the units. β_2 is a measure of the kurtosis or peakedness of a distribution. It is evident that extreme deviations from the mean are given more importance in μ_4, which is based upon their fourth powers, than in μ_2, which is based upon their second powers. For distributions with the same second moment μ_4, and hence β_2, will be greater for the one which has the most extreme values, or the longest tails. But to have its second moment no larger than that of the other distribution, the distribution with longer tails must also have a greater concentration of values around the mean, which will have very small deviations. This will make the distribution with a higher β_2 more peaked. Peakedness or kurtosis is measured relative to the peakedness in a normal distribution, for which the value of β_2 is three. The normal curve is said to be mesokurtic (medi-
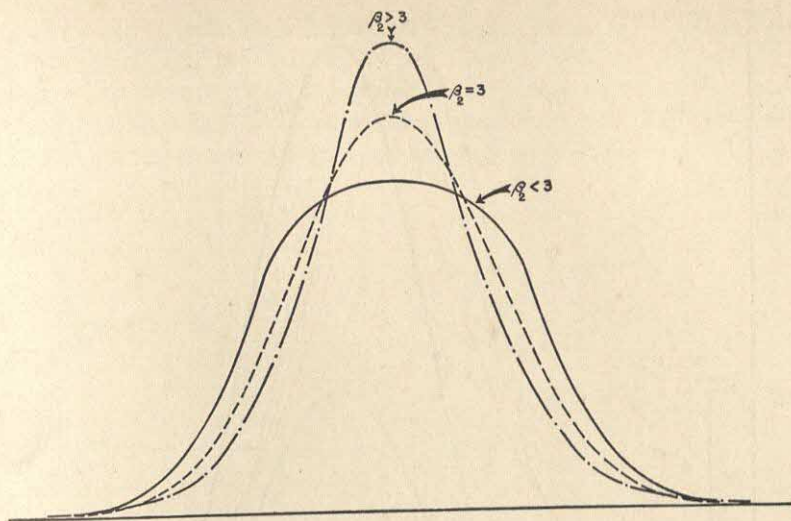
Figure 31. Curves with Different Kurtosis

umly peaked), while a curve showing more peakedness than the normal, with $\beta_2 > 3$, is said to be leptokurtic, and a curve flatter topped or less peaked than the normal, with $\beta_2 < 3$, is said to be platykurtic. Figure 31 shows these three types of curves representing distributions with different values of β_2. Differences in kurtosis are not to be confused with the differences in appearance of normal curves which have different ratios between their horizontal and vertical scales. For the choice of the ratio between the horizontal length representing one standard deviation unit and the vertical unit representing some proportion of cases is entirely arbitrary. Figure 32 shows three normal curves with different horizontal scale-vertical scale ratios. By definition these curves have exactly the same form as measured by a β_1 of zero and a β_2 of 3. They appear different solely because there is a different relation between their horizontal and vertical scales.

There are moments of distributions higher than the third and fourth, but they are used only in advanced theoretical treatments of curves, and we shall not consider them here except to give a general formula defining the nth or any moment about the mean,

$$\mu_n = \frac{\Sigma(X - \bar{X})^n}{N} = \frac{\Sigma x^n}{N} \tag{13}$$

Computation of summarizing measures of form: methods for grouped data. Practical computation procedures for obtaining the second, third and fourth moments, and from them the two beta coefficients, will now be presented and illustrated by computations of them for the distribution
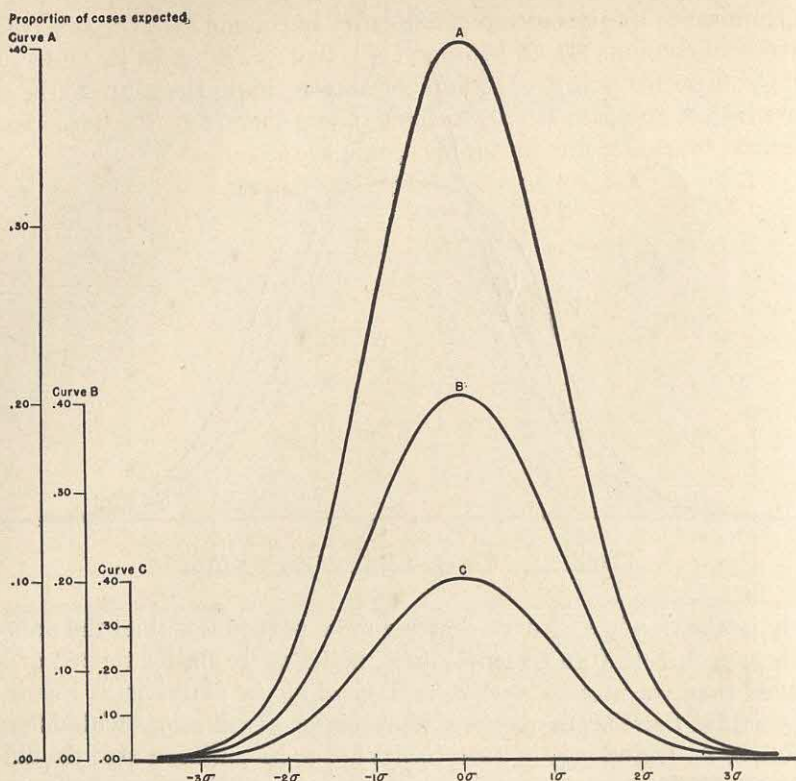
Figure 32.   Normal Curves with Different Ratios between Vertical and Horizontal Scales.

of Table 26. The procedures for obtaining moments and the other summarizing measures based upon them will be illustrated first for grouped data. The grouped data methods are the ones usually employed for obtaining the higher moments. They are used because the features of form the higher moments describe are not often investigated unless the number of cases is fairly large, 100 or more, and if the number of cases is 100 or more, the methods for ungrouped data become quite laborious. The reason that higher moments are not frequently computed for distributions observed among only a small number of cases is that they are summarizing measures less stable than the mean and standard deviation. The explanation of the adjective "stable" used in this sense will be explained in Chapter 16.

Let us examine now columns (6) and (7) of Table 26. Column (6) is obtained by multiplying each entry in column (5) by the corresponding entry in column (3), and column (7) is obtained by multiplying each entry

in column (6) by the corresponding entry in column (3). The sums of the entries of columns (4), (5), (6), and (7), that is $\Sigma fd'$, $\Sigma f(d')^2$, $\Sigma f(d')^3$, and $\Sigma f(d')^4$, together with N, the number of cases, and i the width of the class interval are the data from which we obtain moments. The formulas for the first four moments for grouped data are as follows:

$$\mu_1 = \frac{\left[\Sigma fd' - \dfrac{N\Sigma fd'}{N}\right]i}{N} = 0 \tag{14}$$

$$\mu_2 = \frac{\left[\Sigma f(d')^2 - \dfrac{(\Sigma fd')^2}{N}\right]i^2}{N} \tag{15}$$

$$\mu_3 = \frac{\left[\Sigma f(d')^3 - \dfrac{3\Sigma f(d')\,\Sigma f(d')^2}{N} + \dfrac{2(\Sigma fd')^3}{N^2}\right]i^3}{N} \tag{16}$$

$$\mu_4 = \frac{\left[\Sigma f(d')^4 - \dfrac{4\Sigma f(d')\,\Sigma f(d')^3}{N} + \dfrac{6(\Sigma fd')^2\,\Sigma f(d')^2}{N^2} - \dfrac{3(\Sigma fd')^4}{N^3}\right]i^4}{N} \tag{17}$$

Let us evaluate formulas (15)–(17) with the data from Table 26.

$$\mu_2 = \frac{\left[169 - \dfrac{(-19)^2}{99}\right](10)^2}{99} = 167.02377$$

$$\mu_3 = \frac{\left[131 - \dfrac{3(-19)(169)}{99} + \dfrac{2(-19)^3}{(99)^2}\right](10)^3}{99} = 2{,}299.02222$$

$$\mu_4 = \frac{\left[1{,}405 - \dfrac{4(-19)(131)}{99} + \dfrac{6(-19)^2(169)}{(99)^2} - \dfrac{3(-19)^4}{(99)^3}\right](10)^4}{99}$$

$$= 155{,}809.2222$$

Effects of the grouping assumption on moments. Before we compute betas from the second, third, and fourth moments, let us look more closely into the sort of errors involved in the assumption basic to all treatments of grouped data. It will be recalled that this assumption is that all observed measures falling within an interval may be treated *as if* they all had as their value the midvalue of that class interval. Since we may consider the mean as defined by the first moment—the mean is the value for which the first moment is zero—let us take up the moments one at a time and see how they may be affected by the grouping assumption.

In a normal distribution, or in a distribution approaching normal, the measures which lie above the mean will have their values *increased* by the grouping assumption. For in any class interval above the mean there will usually be more measures in the lower part of the interval than in the upper part of the interval, since generally the density of observed values decreases as the distance of the values from the mean increases. For the same reason the measures which lie below the mean will have their value *decreased* by the grouping assumption. In both cases the grouping assumption makes the measures show up as farther away from the mean than they actually are. For the mean itself, however, (or the first moment) this does not lead to any biased error, because the increases in values of measures lying above the mean will tend to be offset by the decreases in values of measures lying below the mean. Therefore, we do not correct the first moment or the formula based upon it for obtaining the mean from grouped data. Irregularities in points of concentration may cause the mean computed from grouped data to be considerably different from the mean computed from ungrouped data, but we cannot for all distributions predict the direction of the difference. Since the error is not regularly in the same direction, we called this an *unbiased* error.

The total variation, the variance (second moment), and the standard deviation, however, when computed from grouped data will tend to be too large. As we have explained, grouping exaggerates the deviations from the mean, and while in the first moment the positive exaggerations cancel out the negative exaggerations, in the second moment the deviations are squared before summing. Since their squares are all positive, the positive and negative exaggerations cannot cancel each other out. Therefore, formula (15) is a biased formula, giving a value which will be too large unless other irregularities counteract the biasing effect of grouping. We shall, therefore, present a correction for the formula shortly.

The third moment, like the first, is based on odd powers of the deviations which have the signs of the original deviations. Therefore, as in the case of the first moment, the exaggerations of the negative deviations in a cubed state will have the opportunity of counteracting the exaggerations of the positive deviations. There is no biased error involved in computing the third moment by formula (16). Yet, let us emphasize that the higher the moment, the greater the effects of whatever irregularities of concentration there may be in the distribution which violate the grouping assumption. Therefore, there may be considerable difference between third moments computed by methods of grouped and ungrouped data, but again we cannot predict the direction of the difference or correct for it because it is as likely to be in one direction as another—that is, it is unbiased.

The fourth moment, like the second, is based on even powers of the

deviations, the signs of which are all positive. Therefore, the positive exaggerations produced in the deviations by the grouping assumption do not have an opportunity to cancel out the negative exaggerations. Thus, we must correct formula (17) since it gives a biased value of the fourth moment.

The corrections for grouping should be applied only when there are no known important irregularities in concentration, when the distribution approximates normality fairly closely, and when the number of cases is great enough to make us believe the anticipated bias will be more important than irregular, unbiased errors (100 or more).

Sheppard's corrections for moments. The corrections are called Sheppard's corrections in honor of the man who developed them. If we denote corrected moments by the same symbol as that for uncorrected moments with a prime added, the formulas for corrected moments are as follows,

$$\mu'_2 = \mu_2 - \frac{i^2}{12} \tag{18}$$

$$\mu'_4 = \mu_4 - \frac{\mu_2}{2}i^2 + \frac{7}{240}i^4 \tag{19}$$

By substitution in formulas (18) and (19) of the data for the example of the distribution of 99 counties of North Carolina by percentage change in population 1940–1950, we obtain corrected moments, thus,

$$\mu'_2 = 167.02377 - \frac{(10)^2}{12} = 158.69044$$

$$\mu'_4 = 155,809.2222 - \frac{167.02377}{2}(10)^2 + \frac{7}{240}(10)^4$$

$$= 147,749.6973$$

Since the standard deviation is the square root of the second moment, we can obtain the "corrected" standard deviation, (the value we have already used in fitting the normal curve), thus, corrected standard deviation $= \sqrt{158.69044} = 12.597$.

Computation of betas. Using the corrected moments and formulas (10) and (12) we compute β_1 and β_2.

$$\beta_1 = \frac{(2299.0222)^2}{(158.69044)^3} = 1.3224$$

$$\beta_2 = \frac{147,749.6973}{(158.69044)^2} = 5.867$$

These measures have little meaning when considered for a single distribution until the student has had experience with their range of values

in a number of distributions. He may begin to acquire the necessary experience by examining carefully Figures 30 and 31.

SUGGESTED READINGS

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chap. 11.

Deming, William Edwards, *Some Notes on Least Squares* (Washington: Department of Agriculture Graduate School, 1938).

Fry, Thornton C., *Probability and Its Engineering Uses* (New York: Van Nostrand, 1928), Chaps. 8 and 9.

Peatman, John Gray, *Descriptive and Sampling Statistics* (New York: Harper, 1947), Chap. 8.

Yule, G. Udny, and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), pp. 181–189.

Nonquantitative Distributions: Sampling Distributions of Proportions

The nature and utility of the methods of this chapter. We have stated that the purpose of inductive statistics is to enable one to make inductions about a universe from information obtained by observations on a sample drawn from that universe. In any research problem the definition of the universe and the method of drawing a sample from it are extremely important matters, but discussion of them will be delayed until Chapter 17. In this and the chapter immediately following, we shall investigate the methods of making inductions about a larger group of units called a "universe" from observations on a fraction of the units called a "sample," *assuming* that the sample has been randomly drawn from a defined existent universe. In these two chapters we shall study the procedures for the "ideal" case and later we shall examine what approximations are involved in the actual practical applications of sampling in sociological research.

Of the steps listed in Chapter 12 involved in making statistical inductions, the last three are to be considered here. They are as follows:

4. Making estimates of universe values of summarizing measures;
5. Computing measures or estimates of measures of the unreliability of these estimates;
6. Interpreting the estimates of summarizing measures and their measures of unreliability more precisely by making statistical tests of hypotheses about the universe.

The methods for carrying out these steps vary for the different summarizing measures and for the different types of characteristics— quantitative or nonquantitative. In this chapter we shall be concerned with the methods of making inductions regarding the distribution of nonquantitative characteristics, although the general treatment of sampling distributions in the next section and the steps outlined for the

testing of statistical hypotheses in the last sections of the chapter are applicable to all types of characteristics. Furthermore, since quantitative distributions may also be expressed as proportions or percentages of the total number which have measures falling in class intervals, the methods relating to the sampling distributions of proportions may be used for quantitative as well as nonquantitative distributions.

SAMPLING DISTRIBUTIONS OF SUMMARIZING MEASURES

Sampling distributions: standard errors.   If a sample is drawn from a universe, it will supply a certain amount of information about that universe. Any summarizing measure, such as a proportion or a mean, computed from the sample gives us an idea of the value of the corresponding summarizing measure of the universe. Yet a second sample of equal size from the same universe would probably have a somewhat different summarizing measure. A third might be still different, as would a fourth, fifth, and so on. If a great number of such samples of equal size were drawn and the same summarizing measure computed for each, these measures would cluster around the universe value of the summarizing measure, some varying farther from it than others. The distribution of the values of the summarizing measure computed from many, many samples of the same size is called the *sampling distribution* of that summarizing measure. The concept of a sampling distribution is central to all inductive statistics and must be clearly understood before the student can proceed further. A sampling distribution is a quantitative distribution and, therefore, can be described as to the aspects of central tendency, dispersion, and form by the summarizing measures explained in Chapters 8, 9 and 14 for observed distributions. For description of central tendency the mean is used, for description of dispersion the standard deviation is used, and for description of form various devices are used. The standard deviation of the sampling distribution of a summarizing measure is known as the *standard error* of that summarizing measure and is usually indicated by a small sigma with the symbol for the summarizing measure attached (see formula (2) of this chapter).

Parameters and statistics; notation.   The universe value of any summarizing measure of the distribution of one or more characteristics is called a parameter; the value of a summarizing measure observed in a sample is called a *statistic*. Obviously, a universe parameter is a fixed and unchanging quantity, whereas a statistic computed from successive samples is a varying magnitude. To distinguish between universe parameters and sample statistics, it is becoming common to denote parameters with Greek letters and statistics with corresponding Roman letters. For instance, the standard deviation of a sample is denoted by s, while the standard deviation of a universe is denoted by σ. Sometimes a statistic

derived from a sample is used as an estimate of the universe parameter; in other cases the statistic may be slightly modified before being used as an estimate of the universe parameter. In either case, when we do not know the value of the universe parameter but are using an estimate of it based upon sample observations, we indicate the fact that the summarizing measure is estimated by using a circumflex over the Greek letter indicating a parameter. Thus, $\hat{\sigma}$ means an estimate of the universe value of the standard deviation which is based upon information from a sample.

How fortunate it would be for students of statistics if all writers used the same notation, or even if single writers were both consistent and logical in notation throughout a text! Yet even the latter is impossible if one wishes to keep as much as possible of the conventionally accepted notation. One of the difficulties is the fact that certain Greek letters are identical with the corresponding Roman ones, and another difficulty is that certain Greek letters had been adopted for summarizing measures of samples long before the distinction between parameters and statistics was so clearly recognized and honored with a differentiation in notation. Another difficulty is that there are no Greek letters corresponding to certain Roman letters already in use as symbols for summarizing measures. For instance, in proportions p is used to represent the proportion of occurrences of some attribute or event and q to represent the proportion of nonoccurrences. But there is no Greek letter corresponding to q, and it is hence impossible to be consistent in using Greek letters for parameters without changing the conventional letters to entirely different ones.

These difficulties usually result in a series of compromises, differing sufficiently with individual writers to make almost no two systems of notation identical in every respect. The principles by which notation for summarizing measures has been adopted in this book are as follows.

1. If possible, a Roman letter is used for a sample statistic, the corresponding Greek letter for the corresponding universe parameter, and a Greek letter with a circumflex over it for an estimate of the parameter.

2. When the above principle of selection of notation cannot be followed, the same Roman letter will be used for both sample statistic and universe parameter with subscripts identifying which is indicated. In this case the circumflex will still be used over the universe symbol to denote an estimate of it.

The above principles of notation refer only to summarizing measures. A frequency observed in a certain category is regarded as original data, not as a summarizing measure. Therefore, the use as explained in Chapter 7 of (A) to represent the number of individuals possessing attribute A and (α) to represent the number of individuals not possessing attribute A is not governed by the above principles of selection of notation for

summarizing measures. The actual reason for choosing this notation for frequencies in categories of dichotomous nonquantitative characteristics is to familiarize the student with the notation so that he will be able to follow easily the treatment of total and partial association of attributes in Chapter 21, based on G. Udny Yule's fuller treatment.[1]

THE SAMPLING DISTRIBUTION OF A PROPORTION WHEN N IS LARGE

Use of proportions. Since in descriptive statistics we treated the methods appropriate for describing the distribution of a nonquantitative characteristic before those appropriate for a quantitative characteristic, we shall follow the same procedure in inductive statistics, although the reverse order is more customary. Since percentage of occurrence of an attribute can always be expressed as proportion of occurrence, we shall deal only with formulas calling for proportions. These are appropriate, however, for any percentage which can be expressed as a proportion, that is, for any component percentage. The central problem of this chapter is to show how to estimate the proportion of occurrence of an attribute in a universe and how to test various hypotheses about the universe when we have information on only a sample of the universe. For such tests we need to know the description of the sampling distribution of proportions observed in samples of equal size.

The case where the universe parameter is known. Although our aim is to make inferences about the universe from knowledge of the sample— that is, from the particular to the general as is the procedure in all induction—we shall have to consider first the case where the universe parameter is known and deduce the sampling distribution of the statistic. In a universe of unlimited size, or of finite size but so large that the number of units in a sample is negligible compared with the number in the universe, let us define p_u and q_u, thus,

p_u = proportion of units possessing an attribute A

and q_u = proportion of units not possessing an attribute A

Since the sum of the proportion possessing the attribute A and the proportion not possessing the attribute A must be equal to one, we have

$$p_u + q_u = 1 \tag{1}$$

With regard to an attribute,[2] the distribution of that attribute in the universe is completely specified when p_u is known. A measure of dispersion such as a standard deviation is not needed because the individuals are

---

[1] G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), Chaps. 1–3.

[2] We use "attribute" to mean the major category of a dichotomous nonquantitative characteristic. For instance, the distribution of the dichotomous nonquantitative characteristic sex may be described by considering the distribution of the attribute "maleness."

simply divided into two groups, one containing the proportion p_u of all the individuals (this group possessing the attribute), the other containing the proportion q_u of all the individuals (this group not possessing the attribute). Then if the universe parameter p_u is known, q_u is also known from relation (1), and the description of the distribution in the universe is completely specified.

When random samples are drawn from a universe, we shall use the symbols p_s to denote the proportion of the individuals in the sample possessing attribute A. That is, in samples of size N,

$$p_s = \frac{(A)}{N}$$

$$q_s = \frac{(\alpha)}{N}$$

and

$$p_s + q_s = 1$$

where

(A) = number of individuals possessing attribute A
(α) = number of individuals not possessing attribute A
N = number of individuals in a sample

If successive samples of size N are drawn from a universe in which the proportion p_u of individuals possess attribute A, we should expect the proportion of the N individuals in each sample which possess the attribute A to be somewhere near p_u, but not necessarily exactly p_u because of sampling variation. This fits in with the common sense notion that a representative sample will resemble its parent universe to some extent in distributions of characteristics, but will not duplicate its feature exactly. If no factors are operating to make the sample different from its parent universe in the proportion of individuals possessing the attribute except those unexplainable and probably numerous factors determining which units will be drawn in a purely random sample, we label the operating factors "chance." Thus, the phrases "chance variation" or "fluctuations due to chance" or "sampling variation" refer to the differences in distributions or their summarizing measures observable in different samples of the same size which have been randomly drawn from the same universe. It has been established both theoretically and empirically that the larger the size of the random samples, the smaller the differences will be between the summarizing measures of the distributions of their characteristics.

Mean, standard deviation, and form of the sampling distribution of a proportion. For any given universe proportion p_u and size of sample N the distribution of the proportion which may be expected to be observed in the samples, that is, the distribution of the p_s's, has been discovered. The distribution, which we shall call the sampling distribution of a

proportion, will have as its mean p_u, the universe proportion. The standard deviation of this distribution, to which we shall give the notation σ_{p_s} will be,

$$\sigma_{p_s} = \sqrt{\frac{p_u \, q_u}{N}} \tag{2}$$

The form of the sampling distribution will approximate that of the normal curve if N is not too small, and if neither p nor q is too small. Because the form is approximately normal, the table of areas under the normal curve can be used to predict what proportion of samples will have a p_s lying within any designated range of the universe proportion p_u.

Illustration of a sampling distribution of a proportion. Let us illustrate such a distribution with an actual example. Suppose that the proportion of males is .6 in a population (or universe) so large that for practical purposes we can consider it infinite. We draw random samples, each of size 100, and find the proportion of males in each sample. The third column of Table 28 shows the distribution of 100 such samples.

Table 28. DISTRIBUTION OF PROPORTION OF MALES IN 100 SAMPLES OF SIZE 100 DRAWN FROM A POPULATION IN WHICH THE PROPORTION MALE IS 0.60 WITH COMPUTATIONS FOR MEAN AND STANDARD DEVIATION OF DISTRIBUTION.

Proportion male (1)	m (2)	f (3)	d' (4)	fd' (5)	$f(d')^2$ (6)
.47–.48	.475	1	−6	−6	36
.49–.50	.495	2	−5	−10	50
.51–.52	.515	3	−4	−12	48
.53–.54	.535	6	−3	−18	54
.55–.56	.555	7	−2	−14	28
.57–.58	.575	15	−1	−15	15
.59–.60	.595	20	0	0	0
.61–.62	.615	16	1	16	16
.63–.64	.635	12	2	24	48
.65–.66	.655	7	3	21	63
.67–.68	.675	7	4	28	112
.69–.70	.695	3	5	15	75
.71–.72	.715	1	6	6	36
Sums		100		35	581

$$\bar{X} = \bar{X}' + \frac{\Sigma fd'}{N} \, i = .595 + \frac{35}{100}(.02) = .602$$

$$s = i\sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} = .02\sqrt{\frac{581}{100} - \left(\frac{35}{100}\right)^2} = .048$$

While 100 is not a large number of samples, we shall still expect the mean of this distribution of samples to be reasonably close to the universe proportion of .60. If we had a very large number of samples, we would expect the standard deviation of their distribution to be given by formula (2) above. For a sample size of 100 and a universe proportion of .60, formula (2) gives

$$\sigma_{p_s} = \sqrt{\frac{(.6)(.4)}{100}} = .04899$$

Table 28 shows that the mean of our distribution of samples is .602 and the standard deviation is .048. Considering the relatively small number of samples used this is relatively close to the theoretically expected values.

The form of our sampling distribution is approximately normal as is shown in Figure 33. From Appendix Table C we can determine what percentage of the sample proportions would lie beyond various ranges. For instance, we should expect only about 5 percent of the samples to show proportions greater than the mean plus two standard deviation units,

$$.60 + 2(.049) = .698$$

or smaller than the mean minus two standard deviation units,

$$.60 - 2(.049) = .502$$

In the distribution of Table 28 we find 4 percent of our cases lying outside these limits. This is reasonably close considering the fact that we only have 100 samples here.

Illustration of the concept of "probability." Statisticians use the concept of probability to interpret the proportions of areas under a curve in the above situation. A probability is a fraction, usually expressed as a decimal fraction, although not necessarily so, with limiting values of zero and one. A probability is sometimes defined as the ratio of all "favorable" events to all possible events; therefore, it may be considered a proportion. If in the above sampling distribution we consider for the moment a sample proportion's falling inside the range $p_u \pm 2\sigma$ as a favorable event, and a sample proportion's falling anywhere between zero and one as a possible event, we can say that the probability of any single sample proportion's falling inside the range .502 — .698 is .95. Or we can say that the probability that a sample proportion will be as unusual as .698 or more unusual than .698 is .05. We shall not attempt to present a systematic treatment of probability in this text but rather to introduce the student to the concept and its uses through practical illustrations.

If, as in the hypothetical case above, we know the population pa-
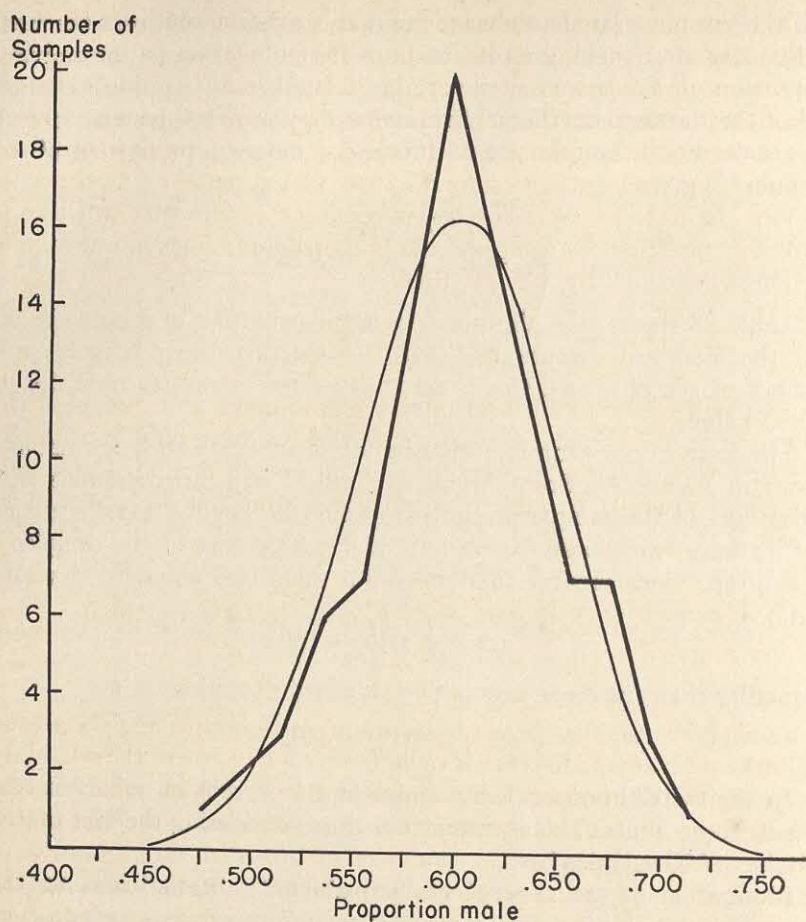
Figure 33. Distribution of Percent Male in 100 Samples of Size 100 from a Universe with $p_u = 0.60$ and the Theoretically Expected Normal Distribution. (Source: Table 28.)

rameter, p_u, the predicting of the proportions of samples which will be expected to fall within certain ranges or the stating of the probability that they will fall within certain ranges is a straightforward matter. This is not usually the case in sociological research, however. The problem we have to face in practice is more difficult—that of inferring information about the population parameter when we know only one sample statistic. For it is obvious that if information were available for the complete universe, there would not be much reason for making a sampling study to infer information about the universe.

Estimation of the parameter. There are several ways in which we can go about inferring information concerning the population parameter

from the sample statistic. Developments in statistical methods have led to the method of making an estimate of the population parameter and then setting up confidence limits around this estimate. To illustrate this method, let us suppose that in a random sample of 100 persons from a universe in which the proportion of males is unknown, we have observed 60 males. The problem is to infer what we can about the proportion of males in the universe by using the information obtained from this one sample of 100. First, we compute the proportion of males in our sample by dividing 60, (A), by 100, N, thus,

$$p_s = \frac{60}{100} = .6$$

The first half of the problem of inference is to make an estimate of the proportion of males in the universe, and when we have only one sample, the best estimate of the universe proportion we can make is the sample proportion observed. In the customary notation this is expressed as follows,

$$\hat{p}_u = p_s \tag{3}$$

Evaluation of (3) with the observed p_s gives us as our estimate of the universe proportion,

$$\hat{p}_u = .6$$

We know, however, that p_s varies from sample to sample and the second half of the problem of inference is to determine a measure of the reliability of this estimate. Formerly one computed the "probable error" of the estimate, but the older interpretation associated with the use of the probable error is open to the criticism of incorrect interpretation of probability. The way more generally preferred now for handling the matter of reliability is to set up confidence limits for the estimate. The explanation and interpretation of confidence limits is somewhat cumbersome at first, but once they are fully understood, the lengthy verbal interpretation of them can be omitted, since it can now be assumed that other statisticians know what they mean.

Confidence limits: definition and interpretation. Confidence limits are two points, one above and one below an estimate, which we can determine and expect to be right 95 times out of 100 (or any other proportion of times we choose) in saying that they include the universe parameter. Notice that we are *not* interpreting confidence limits as the points delimiting a range within which the probability is .95 that the universe parameter falls. We cannot do that for the simple reason that the universe parameter is a fixed value (even if we do not know it), and it either lies within the range marked off by the confidence limits or it does not. The ordinary concept of probability is not even useful in such a situation, for the

probability of an event is defined as the ratio of favorable occurrences to all occurrences in the universe of events from which the particular event may be considered randomly drawn, and thus is a fraction between zero and one indicating the degree of expectation of an event. Therefore, for any one single event which has already happened, the probability of its having happened a certain way is either one or zero according to whether it did or did not happen that way.

In the correct interpretation of confidence limits we used the phrase, "would be correct 95 out of 100 times," which involves the notion of the operation's being repeated. The phrase actually means that if we were to set up such 95-percent confidence limits successively for estimates made from one sample after another for an infinite number of times, in 95 out of 100 times we should enclose the population proportion within the confidence limits. We cannot, of course, be absolutely sure that we have enclosed the population parameter in this particular case, for it might be one of the five times in 100 where we should expect to fail. If we wish to have more confidence that we have enclosed the population mean by the confidence limits, we can use instead of the 95-percent confidence limits 99-percent confidence limits, in which case we should expect to be correct in our statement about enclosing the population proportion in 99 times out of 100. In the matter of interpretation, let us emphasize that the probability of .95 or .99 implied in the level of "confidence" has to refer to something which is not fixed. Since the universe parameter is a fixed and unchanging magnitude, the interpretation of confidence limits should *never* involve the statement that the probability of its falling within an interval is some value other than zero or one; rather we must interpret the probability of .95 or .99 to refer to a set of events (such as our making successive statements) some of which may possess one attribute (such as being correct) and all the rest of which possess the opposite attribute (such as being wrong). Then the notion of repeated setting up of confidence limits is necessary for the interpretation of a single pair, since the probability of the correctness of our statement has meaning only when this statement is thought of as one of many such statements, of which the proportion of correct ones is .95, and therefore the probability of a single statement's being correct is .95.

Approximate procedure for obtaining confidence limits. The actual procedure for obtaining 95-percent confidence limits is not difficult. We shall explain first an approximate procedure and then a more exact one. The problem is to locate two values, p_1 below p_s and p_2 above p_s, which are called the confidence limits of the estimate $\hat{p}_u = p_s$ and which enclose the confidence range or confidence interval of the estimate. We shall explain first how to find the value of p_1. We can think of the problem of determining the value of p_1 as the problem of finding what distance below

p_s it lies on a scale of all possible values a proportion can take such as that shown in Figure 34. The scale, of course, extends from zero to one and we know the position of the observed $p_s = .6$ on the scale. The point of p_1 must be located on this scale at a certain distance below p_s. This distance must be such that if p_1 were the population parameter, the probability of observing a sample proportion as unusual as p_s would be
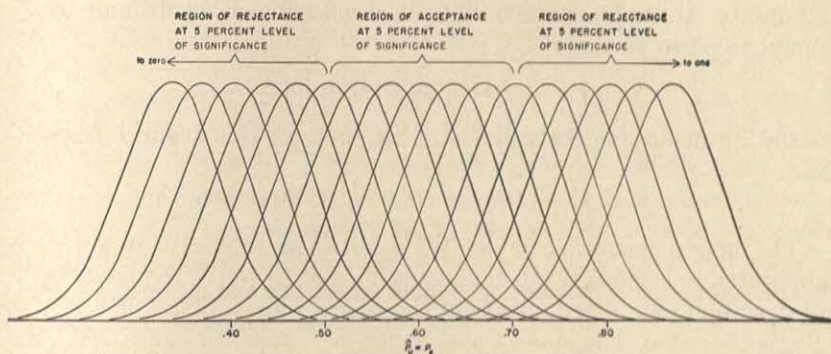
Figure 34.   Tests of Hypotheses Implied by 95-percent Confidence Limits.

.05. To find the distance we need to know the standard deviation of the sampling distribution of samples of size N drawn from a universe with a proportion equal to p_1. The standard deviation of such a distribution is, of course, the standard error of a proportion with a value of p_1. We could obtain the desired standard error by substituting the value of p_1 in formula (2) if we knew in advance the value of p_1, but that is what we are trying to find. Since the standard error of a proportion in a sample of 100 cases does not change greatly as the value of the proportion changes, we shall use in formula (2) the value of p_s as an approximation to p_1. Thus, we get an approximation to the standard deviation unit we want by substituting $p_s = .6$ in formula (2),

$$\text{Approx. } \sigma_{p_1} = \sqrt{\frac{p_s q_s}{N}} = \sqrt{\frac{(.6)(.4)}{100}} = .049$$

We then turn to Appendix Table A of areas under the normal curve to find what number of standard deviation units includes exactly 95 percent of the area. Since the table refers to the area on only one side of the mean, we shall have to find from it what number of standard deviation units includes one half of 95 percent of the area—that is, 47.5 percent of the area. This time we look at the cell entries to find the one at or nearest to

4750 and read the table backwards to see what number of standard deviation units corresponds to this proportion of the area. We find that it is 1.96 standard deviation units which include 47.5 percent of the area on one side, which means that between -1.96 standard deviation units and $+1.96$ standard deviation units, 95 percent of the total area is included.

We have now found the approximate size of the standard deviation unit and the multiple of a standard deviation required for the given probability. Multiplying one by the other will tell us what distance away from p_s to place p_1, thus,

$$p_s - p_1 = 1.96(.049) = .096$$

Finally, to determine the value of p_1 we subtract this distance from p_s, thus,

$$p_1 = p_s - .096 = .6000 - .0392 = .504$$

The upper confidence limit, p_2, is obtained similarly. Again p_s is substituted in formula (2) to get an approximation to the required standard deviation unit. The multiple of the standard deviation unit required, 1.96, is the same. Therefore, the distance $p_2 - p_s$ is the same, .096, and p_2 is actually located by adding this distance to p_s, thus,

$$p_2 = p_s + .096 = .600 + .096 = .696$$

In presenting these results to others we should say that we estimate the proportion of males in the universe to be .60, with 95-percent confidence limits (rounded to two places) of .50 and .70, or with a 95-percent confidence range of from .50 to .70.

The above computation of confidence limits is not exactly correct because of the fact that the standard deviation of the sampling distribution of a proportion changes in size with changes in p_u. In spite of this we have used standard deviation units computed as if the universe proportion were .60 to measure off distances of sampling distributions from universes with proportions of .50 and .70. If N is moderate or large and if the proportion is somewhere near the .50 level, such an approximation is accurate enough.

Since the approximate procedures for determining confidence limits are the ones most frequently used, they are condensed into formulas below for convenient reference. For 95-percent confidence limits,

$$\text{Approx. } p_1 = p_s - 1.96\sigma_{p_s} \tag{4}$$
$$\text{Approx. } p_2 = p_s + 1.96\sigma_{p_s} \tag{5}$$

where the symbols have meanings as defined above. By reference to Appendix Table A we can find that the multiple of a standard deviation unit required to cut off .005 of the area at one end or .01 of the area on

both ends is 2.58. Therefore, formulas for the 99-percent confidence limits can be obtained by substituting 2.58 for 1.96 in (4) and (5). Thus, for 99-percent confidence limits,

$$\text{Approx. } p_1 = p_s - 2.58\sigma_{p_s} \tag{4A}$$

$$\text{Approx. } p_2 = p_s + 2.58\sigma_{p_s} \tag{5A}$$

More accurate procedure for obtaining confidence limits. The more accurate procedure for obtaining confidence limits consists of solving the following equation which simply expresses the relation that the distance between the sample proportion (which is the estimate of the universe proportion) and its lower 95-percent confidence limit is equal to 1.96 times the standard error of the lower confidence limit. The equation is

$$p_s - p_1 = 1.96 \sqrt{\frac{p_1 q_1}{N}} \tag{6}$$

After substituting the numerical value of our example for p_s and N, and $1 - p_1$ for q_1, we have

$$.6 - p_1 = 1.96 \sqrt{\frac{p_1(1 - p_1)}{100}}$$

We solve this quadratic equation by squaring both sides, collecting terms, reducing to the standard form of a quadratic, and using the formula for the solution of a quadratic. Since every quadratic equation has two roots, we get two values for p_1 which satisfy the original equation. For our example these values are .691 and .502. Actually, one of these is the lower confidence limit, and the other is the upper confidence limit. We will get the same two values if we set up an equation similar to (6) for p_2.

Comparison of confidence limits obtained by the two procedures. These values for the confidence limits, .502 and .691, do not differ appreciably from those found by the approximate procedure given at first, .504 and .697. It is interesting, however, to note the direction of their differences and to see why the differences occur in that direction. Both the upper and lower confidence limits are slightly lower when computed by the more accurate procedure. This is explained as follows. The expression for the standard error of a proportion,

$$\sigma_{p_s} = \sqrt{\frac{p_u q_u}{N}} \tag{2}$$

has its maximum value for any N when $p = q = .5$. Therefore, standard deviation units of the sampling distribution of the proportion of any size of sample are greater the nearer the proportion is to .5. When by solving the quadratic equation for the lower limit, we used .504 as the p_u around

which the standard deviation units were computed, we obtained slightly larger standard deviation units than when in the approximate method we used .600 as the p_u. Then 1.96 of these larger standard deviation units locate a point farther away from .600 than the same multiple of the smaller standard deviation units, and the p_1 located is consequently lower in value. Similarly, the standard deviation units computed with .691 as p_u are slightly smaller than those computed with .600 as p_u; 1.96 of these smaller units locate a point nearer to .600 than the same multiple of the larger standard deviation units, and consequently the p_2 located is lower in value. When the number of cases is large, the difference is negligible since we would probably not record the confidence limits to more than two or three decimal places anyway; but when the number of cases is small, it must be remembered that the confidence limits located by the approximate method will always be biased in the direction of being too far away from the .5 point.

Assumptions involved in both procedures for obtaining confidence limits. Now the procedure we have called "more exact" is still an approximation based upon the assumption that the sampling distribution of p_s is continuous and normally distributed. Since the number of units possessing an attribute A in any sample is a whole number, the distribution of sample proportions is not continuous. In a sample of size N, p_s can take only the following values:

$$\frac{0}{N}, \frac{1}{N}, \frac{2}{N}, \frac{3}{N}, \ldots, \frac{N}{N}$$

It can be seen that the greater N is, the greater number of values p_s can take and the nearer its distribution approaches continuity. In fact, the distribution of p_s is of the same type as that of "fertility ratio," a type III A characteristic which, when N is great enough, we treat as a type III B characteristic because it appears to be continuous for all practical purposes. On the other hand, if p_u is very small or very large, near to zero or to one, the range of the possible values of p_s listed above which would be observed would be very greatly restricted, and the distribution again would not approach normality closely since the restricted range on one side would produce a skew. It is not possible to set definite limits for either N or p_u alone, within which the double approximation of continuity and normality is close enough to be used. There is an empirical rule of thumb, however, based on the values of both N and $\hat{p}_u$, which can be used as a guide. It is evident that in the formula for the standard deviation of the sampling distribution of a proportion p_u and q_u are interchangeable. Because of this, when we set up confidence limits for $\hat{p}_u$, we can obtain corresponding confidence limits for $\hat{q}_u$ by the two subtractions,

$$q_1 = 1 - p_2 \tag{7}$$

$$q_2 = 1 - p_1 \tag{8}$$

This is reasonable because q_s, the proportion of N units of a sample which does not possess the attribute A is always equal to $1 - p_s$, when p_s is the proportion which does possess the attribute A. Therefore, in designating the proportions it is quite possible to take p_s always to be the smaller proportion. If we do this, an empirical rule [3] is that the sampling distribution of a proportion may be considered as normal and continuous if

$$Np_s + 9p_s > 9, \text{ when } p_s < q_s \tag{9}$$

Let us apply this rule to our example to see if we are justified in assuming continuity and normality. Since .6 is less than .4, we shall have to consider p_s as .4, the smaller of the two complementary proportions.

$$(100)(.4) + 9(.4) = 43.6 > 9$$

Therefore, we see that we are safe in using the normal curve as an approximate description of the sampling distribution. Suppose that the number of cases in the sample had been only 6, then

$$(6)(.4) + 9(.4) = 6.0 < 9$$

and we see that we could not have used the normal curve to describe the sampling distribution. In fact, we could not have observed a proportion of .4 in a sample of 6, for the only possible values p_s can take in such a case are the following:

$$\frac{0}{6}, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}$$

On the other hand, if p_s had been very small, say .001, even if N were 600, we would get

$$(600)(.001) + 9(.001) = .609 < 9$$

and we see that we are not justified in assuming that the sampling distribution is continuous and normal.

Special methods have been developed for handling the case where N is quite small and the case where p_s is quite small. The latter case is handled by the methods of a Poisson distribution and will not be treated in this text. The former case, when N is small, is handled by the methods of the binomial distribution, which we shall now examine.

---

[3] See Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, 1949), p. 58.

THE SAMPLING DISTRIBUTION OF A PROPORTION WHEN N IS SMALL

The binomial distribution. A review of a few simple algebraic principles is necessary before proceeding with the statistical applications of the binomial distribution. Any student should remember from his high school algebra that

$$(q + p)^2 = q^2 + 2qp + p^2$$
$$(q + p)^3 = q^3 + 3q^2p + 3qp^2 + p^3$$

and so on for higher powers. Although he has probably forgotten it by now, even in high school algebra the student probably learned also the general rule for the expansion of a binomial to any integral power. This can be given by formula, thus,

$$(q + p)^N = q^N + \frac{N}{1} q^{N-1} p + \frac{N(N - 1)}{2 \cdot 1} q^{N-2} p^2 + \cdots$$
$$+ \frac{N(N - 1)(N - 2) \cdots (N - r + 2)}{(r - 1)!} q^{N-r+1} p^{r-1} + \cdots$$
$$+ \frac{N(N - 1)(N - 2) \cdots (2)(1)}{(N - 1)!} qp^{N-1} + p^N \qquad (10)$$

The application of the binomial expansion to the description of the sampling distribution of a proportion is made as follows. First we let the symbols in (15) have the meanings specified below

p = proportion of individuals in the universe possessing attribute A, that is, p_u

q = proportion of individuals in the universe not possessing attribute A, that is q_u

N = number of individuals in each of the random samples

Then the successive terms of (10) evaluated with the appropriate values of p, q, and N give a set of numbers proportional to the frequencies with which we should expect to observe the various possible values of p_s in the order

$$\frac{0}{N}, \frac{1}{N}, \frac{2}{N}, \frac{3}{N}, \cdots, \frac{N}{N}$$

Use of the binomial distribution where the universe proportion is known. We shall illustrate this application of the binomial expansion with the example we mentioned which could not be treated by the methods assuming a normal distribution. We shall assume that the universe parameters are known and are

$$p_u = .6$$
$$q_u = .4$$

We wish to discover what is the sampling distribution of p_u or p_s from samples of size 6.

Let us first substitute $N = 6$ in expression (10).

$$(q + p)^6 = q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 \\ + 15q^2p^4 + 6qp^5 + p^6$$

There will always be $N + 1$ terms in the expansion of a binomial raised to the Nth power. These terms correspond to the $N + 1$ possible values which p_s (or $\hat{p}_u$) can take, as shown in Table 29. Let us notice certain

Table 29. TERMS OF THE BINOMIAL EXPANSION FOR SAMPLES OF SIZE SIX WITH THE CORRESPONDING NUMBER AND PROPORTION OF A'S OBSERVED IN THE SAMPLE

Term of binomial	Number of A's observed in sample	Proportion of A's observed in sample, represented by the term, p_s
q^6	0	$\dfrac{0}{N}$ $= .0000$
$6q^5p$	1	$\dfrac{1}{N} = \dfrac{1}{6} = .1667$
$15q^4p^2$	2	$\dfrac{2}{N} = \dfrac{2}{6} = .3333$
$20q^3p^3$	3	$\dfrac{3}{N} = \dfrac{3}{6} = .5000$
$15q^2p^4$	4	$\dfrac{4}{N} = \dfrac{4}{6} = .6667$
$6q\,p^5$	5	$\dfrac{5}{N} = \dfrac{5}{6} = .8333$
p^6	6	$\dfrac{6}{N} = \dfrac{6}{6} = 1.0000$

Source: Formula (10).

features of the table which are true for every such expansion. First, there are $N + 1$ (seven when N is six) terms of the binomial expansion which correspond to the $N + 1$ possible values of p_s. The exponent of p in each term tells the number of observed A's in the sample, while the exponent of q tells the number of not-A's observed in the sample. These relations would hold for varying values of p and q. The use of the correspondence

between the terms of the expansion and the values of p_s is as follows—when the terms in the leftmost column of Table 29 are evaluated for a particular p and q they tell us the relative frequency with which the corresponding p_s's would be observed. Thus, they afford a description of the sampling distribution of $\hat{p}_u$ or p_s.

Since $p + q$ is always equal to one by definition and since one raised to any power is always one, the sum of all the terms in the leftmost column of Table 29 when evaluated for a particular p and q must be equal to one. Then if we actually substitute the values of p and q of our particular example in these successive terms of the binomial, we get a series of fractions which add up to one as shown in Table 30. Each one of these fractions

Table 30. DESCRIPTION OF SAMPLING DISTRIBUTION OF p_s WHEN $p_u = .6$, $q_u = .4$, AND $N = 6$

Value of p_s	Proportion of times expected		Number of times expected in 1,000 samples
	Binomial term	Actual value	
(1)	(2)	(3)	(4)
.0000	q^6	$(.4)^6 = .004096$	4
.1667	$6q^5p$	$6(.4)^5(.6) = .036864$	37
.3333	$15q^4p^2$	$15(.4)^4(.6)^2 = .138240$	138
.5000	$20q^3p^3$	$20(.4)^3(.6)^3 = .276480$	276
.6667	$15q^2p^4$	$15(.4)^2(.6)^4 = .311040$	311
.8333	$6q\,p^5$	$6(.4)\,(.6)^5 = .186624$	187
1.0000	p^6	$(.6)^6 = .046656$	47

Source: Table 29.

or proportions represents the proportion of samples we should expect to observe in a series of samples with the corresponding frequency or proportion of A's. For instance, if we drew repeatedly samples of size six from a universe which has .6 of its individuals possessing the attribute A, we should expect to get samples with no individuals possessing the attribute A in about four times out of 1,000 times. We should expect to get samples with only one individual possessing the attribute A in about 37 out of 1,000 times, and so on to the expectation of samples with all individuals possessing the attribute in about 47 out of 1,000 times.

Actually, we would not expect the distribution of 1,000 samples to be identical with the one given as a frequency table in the last column of Table 30, even if the method of obtaining the samples were ideally random. However, the relative frequencies shown in Table 30 are what we should expect an observed distribution to approach if the number of samples were increased indefinitely.

Use of the binomial distribution when the universe proportion is unknown. So far we have been considering the case where the universe proportion was assumed to be known. As we have pointed out previously, this is not the usual case in practical experience, where we are trying to infer information about the universe and its parameters from observations made on a sample. The first step necessary in the practical situation is to estimate the proportion in the universe, for which as before we use the observed proportion in the sample,

$$\hat{p}_u = p_s \tag{3}$$

When the number of cases is quite small, as in the example we have been considering, we know that this estimate is quite unreliable. By solving equations of an order equal to the number of cases in the sample, it is possible to set up confidence limits, but this is too lengthy a procedure for practical use. Therefore, when N is too small to use the normal distribution as an approximation, we usually forego the setting up of confidence limits and instead test directly hypotheses about the unknown universe parameter by means of the binomial distribution we have been describing. The methods of testing certain types of hypotheses will be discussed in the next section of this chapter.

TESTS OF STATISTICAL HYPOTHESES

Limitations of statistical inductions. We are now ready to illustrate with actual examples the procedure of testing a statistical hypothesis, which we have referred to several times. Before we give examples, however, even at the risk of repeating material already presented, let us summarize a few general facts about the testing of any statistical hypothesis.

In statistical induction we are never able to prove anything in an absolute sense. We are never able to say on the basis of information supplied by observation of a sample that this or that is *certainly* true or not true about the universe from which it was drawn. This is the first fact to be kept continually in mind when attempting to make statistical inferences.

The second fact of importance is that we can by statistical methods come nearer to proving that something is *not* true about a universe than that something *is* true. This means that often we shall use a negativistic approach. If we want to establish one hypothesis, we shall not test it directly but shall formulate the opposite hypothesis, which we shall call the null hypothesis, and test it on the basis of the evidence from our sample. If the evidence is such as to cause us to reject or discard the null hypothesis and if the hypothesis we wanted to establish is the *only*

alternative hypothesis, then the rejection of the null hypothesis is the equivalent of the confirmation of the original hypothesis. Often, however, there may be a number of alternative hypotheses, so that no single one of them is confirmed by rejection of the null hypothesis; or if one hypothesis covers all possible alternatives, it will have to be rather broad and non-specific. Therefore, the confirmation of a specific hypothesis is not so frequent as rejection of a specific hypothesis.

This negativistic approach to acquiring knowledge about a universe by formulating null hypotheses and then rejecting them on the basis of evidence seems almost the equivalent of setting up straw men merely to shoot them down. Yet, in so doing, certain logical possibilities are elimi-nated, and the range of the remaining possibilities is narrowed. It is a cautious way of proceeding as are most scientific procedures. Because the results of statistical analysis are often inconclusive, and at best can only reveal the improbability of obtaining results as unusual as those observed if the tested hypothesis is true, such analysis and interpretation are unsatisfying to the person who wants absolute and final answers and positive proof.

The steps involved in the statistical testing of a hypothesis. The exact steps in testing a statistical hypothesis are as follows:

1. Formulation of a hypothesis (often a null hypothesis) about one or more parameters of the universe to be tested;

2. Description of the sampling distribution expected of estimates of the parameter or parameters (in this step we usually have to use estimates of the standard deviation of a sampling distribution based upon information from the sample, since we do not as a rule know the universe parameters);

3. Determination of the probability that a value as unusual as that of the sample statistic would have been observed from this sampling distribution; [4]

4. Rejection or nonrejection of the hypothesis on the basis of the value of the probability determined in 3;

5. Interpretation of the test in terms of the actual problem, both for the information it yields on the tested hypothesis and on alternative hypotheses.

Illustration of a test of a statistical hypothesis when N is large. We shall illustrate these steps in connection with the example we have already mentioned. Suppose the sample of 100 is a sample of people drawn from a population large enough to be considered infinite in comparison with the size of the sample, and that the attribute A is "maleness." Let us also suppose that 60 males are observed in this sample of 100, which makes p_s equal to $\frac{60}{100}$ or .6. Such a value of p_s would make us inclined to believe that we are sampling from a universe which has more males than females

---

[4] We are using the phrase "as unusual as" to mean exactly the same as the more cumber-some phrase, "as unusual as or more unusual than," often used by statisticians.

in it, but how can we be sure that the excess of males observed in the sample cannot be explained by sampling fluctuation? We shall make a statistical test of a hypothesis to see what we are able to say about the universe and with what degree of confidence. The hypothesis that we are interested in establishing is that in the universe there are more males than females, or in symbols,

$$p_u > .5 \tag{11}$$

Our evidence is in accord with this hypothesis, but that does not by any means establish the hypothesis, for our evidence might be in accord with other hypotheses as well. From here let us follow the steps of making a test as outlined above.

1. *Formulation of the hypothesis to be tested.* Since p_u can take values only from zero to one, the null hypothesis, or the hypothesis which covers all possible alternatives to the one we want to establish can be formulated thus,

$$p_u \leq .5 \tag{12}$$

If we can find evidence for rejecting this hypothesis, then hypothesis (11) will be confirmed.[5] Hypothesis (12) is in too general a form to be tested since it covers a range of possible values of p_u from zero through .5. To narrow it to testable dimensions, we reason thus: it is obvious that the observed p_s of .6 would be more likely to be observed from a p_u of .5 than from a p_u of any lower value; therefore, if we test the hypothesis,

$$p_u = .5 \tag{13}$$

and find reason to reject it, there would be even more reason to reject the other possible values of p_u included in hypothesis (12). Then we select hypothesis (13) as the specific null hypothesis which we shall test statistically.

2. *Description of the sampling distribution of estimates of the parameter about which the hypothesis has been formulated.* As long as we are testing a hypothesis that p_u is some definite value, in this case .5, we proceed as if this is the value of the proportion in the universe. Then the sampling distribution of sample proportions observed in samples of size 100 in a universe with a p_u of .5 is what we wish to describe. First, we know that the mean of this distribution will coincide with the universe proportion; that is, it will be equal to .5. We can find the standard deviation of the sampling distribution by evaluating the formula given earlier,

---

[5] Various writers use "confirmed," "sustained," "affirmed," "upheld," and other synonyms in this situation. The student should differentiate between any of these words, which denote one possible verdict appropriate for the *alternative* to the null hypothesis, and the words "accepted" or "not rejected," which denote a verdict appropriate for the null hypothesis.

$$\sigma_{p_s} \text{ or } \sigma_{\hat{p}_u} = \sqrt{\frac{p_u q_u}{N}} = \sqrt{\frac{(.5)(.5)}{100}} = .050 \tag{2}$$

Since

$$Np + 9p = (100)(.4) + 9(.4) = 43.6 > 9$$

we can safely use the normal distribution to approximate the form of the sampling distribution. To summarize then, the sampling distribution has a mean of .5, a standard deviation of .050, and an approximately normal form.

3. *Determination of the probability that a sample proportion as unusual as .6 would be observed in this sampling distribution.* The sample proportion $p_s = .6$ deviates from the mean of the distribution $p_u = .5$ by .1. The problem is to determine the probability that a sample proportion deviating as much as .1 would be observed in the sampling distribution of a proportion with a standard deviation of .050. The deviation must first be expressed in terms of standard deviation units, thus,

$$\frac{.6 - .5}{.050} = \frac{.1}{.050} = 2.0 \text{ standard deviation units}$$

We turn to Appendix Table C and find that the probability of getting a sample as unusual as ours from a universe with a proportion of .50 is .0454. This means that fewer than five out of 100 samples would be expected to show a deviation as great in absolute value as .1 or greater. Usually, we state this in terms of probability and say that the probability is .0454 that a deviation so unusual would be observed from this sampling distribution.

4. *Rejection of the hypothesis.* It is possible that our sample was drawn from a universe with a proportion of .5, but it is unlikely. The probability of .0454 is sufficiently small that we are not willing to base our conclusions on the belief that we happened to get one of the 45 samples out of 1,000 which would be expected to show such an unusual sample proportion. Therefore, we reject hypothesis (13) that the universe proportion is .5, and by an extension of the same reasoning we reject also the broader hypothesis (12) that the universe proportion is equal to or less than .5. We feel that we are justified in shooting down the straw man we set up because of the improbability of having observed what we did *if* the hypothesis were correct. There is a subjective element here in the choice of level of improbability sufficient to justify rejection. Some statisticians reject a hypothesis when the probability is less than .05, others only when it is less than .02, others when it is less than .01 or even .001. Whatever level of probability is chosen, below which probabilities obtained will indicate a verdict of rejection, that level is called a "level of

significance." Discussion of the choice of levels will be treated in Chapter 19.

5. *Interpretation of the test.* The statistical test itself can go no further than indicating that it is highly improbable that the sample was drawn from a universe where $p_u = .5$. But the practical research worker will continue the interpretation in terms of the alternative hypothesis (11) which he wishes to establish, and will conclude that the universe from which the sample was drawn must have a proportion of males greater than .5. This is quite justifiable since the conclusion is a proposition including all possible alternatives to the rejected hypothesis. A word of warning must be given, however. It is *not* justifiable to take the complement of the probability determined above,

$$1 - .0454 = .9546$$

and use the value to describe the probability of the correctness of the hypothesis we are interested in establishing, that $p_u > .5$. There can be no exact mathematical statement of the probability of the correctness of a hypothesis in the type of situation we are dealing with. This is the reason why the statistician has at times to deal in double negatives, apparent circumlocution, and straw men.

We have not narrowed the field of possibilities about the sex composition so much as we might when we have reached the conclusion that there are more males than females in the universe. If we wished to proceed further we might narrow the range about .60 by testing successive hypotheses that $p_u = .52$, $p_u = .54$, etc. and from the other side by such hypotheses as $p_u = .70$, $p_u = .68$, $p_u = .66$, etc. If we did this, and if we set .05 as the level of probability, below which we would reject hypotheses, we should arrive at the 95 percent confidence limits, .50 and .70 determined earlier in the chapter. This is why the statement was made that confidence limits are implicit tests of statistical hypotheses. Figure 34 illustrates graphically the implicit testing of hypotheses by confidence limits, with the proportions carried to only two decimal places. Each normal curve in the figure represents the sampling distribution of proportions in samples of 100 drawn from the universe with its mean at the mean of the sampling distribution. Let us think of the curves and their means as representing a series of possible hypotheses which one might test. The confidence limits tell us that all of the hypotheses describing universes with p_u's below .50 or above .70 would be rejected, if we used a 5-percent level of significance, and that all of the hypotheses describing universes with p_u's between .50 and .70 would be accepted. While the determination of confidence limits makes unnecessary the testing of single hypotheses such as we have been describing, there are other types

of hypotheses of tremendous practical importance which are not tested by confidence limits. They will be considered in Chapter 19.

Illustration of a test of a statistical hypothesis when the N is small. In the case of very small samples where confidence limits are not conveniently computed, the testing of specific hypotheses is a better way to handle the matter. Let us consider again a universe with an unknown proportion p_u of males, from which a sample of six has been observed to have four males, making $p_s = \frac{4}{6} = .6667$. We may wonder if we can conclude from this information that the universe from which the sample was drawn has more males than females in it. Let us again follow the same five steps for testing a statistical hypothesis.

1. *Formulation of the hypothesis.* As before, since we want to establish the hypothesis that $p_u > .5$, we set up the general null hypothesis that $p_u \leq .5$, which can be rejected if the more specific null hypothesis $p_u = .5$ is rejected. Thus, $p_u = .5$ is the hypothesis which we shall test.

2. *Description of the sampling distribution of $\hat{p}_u$.* Since the product

$$6(.5) + 9(.5) = 3 + 4.5 = 7.5 < 9$$

we cannot use the normal curve to describe the distribution but must make use of a binomial expansion. The distribution is described by the expansion of

$$(q + p)^N$$

where

$$p = p_u = .5$$
$$q = q_u = .5$$
$$N = 6$$

Let us first substitute only 6 for N and expand this binomial,

$$q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 + 15q^2p^4 + 6qp^5 + q^6 \tag{19}$$

Now the fact that $p = q = .5$ makes this expansion much more easily evaluated than the previous one where p and q were unequal. Notice that the sum of the exponents of p and q in every term is 6. Because of this and because p and q are identical, the part indicated by the letters in each term of (19) is equal to $(.5)^6$ or to $(\frac{1}{2})^6$ or to $\frac{1}{64}$. The sum of the numerical coefficients of all the terms in (19) is 64, so that the sum of the products of each numerical coefficient times $\frac{1}{64}$ is one, as before. Then the relative frequencies may be computed as shown in Table 31. This table describes the sampling distribution as fully as is necessary, and in this case we shall use the relative frequency (percentage) table as the description of the distribution, rather than the descriptive measures—mean, standard deviation, and description of form—as with larger samples.

3. *Determination of the probability of observing a sample proportion as unusual as .6667 in the sampling distribution.* We have drawn horizontal

lines above and below the mean, $p_u = .5$ of this distribution for emphasis. Now we can read from the two rightmost columns that the probability of observing exactly four males in a sample of six is 15 out of 64 or, expressed as a decimal fraction in the customary way of indicating a probability, .234375. But this probability is not what we want to test our hypothesis. We want to know what is the probability of having observed a p_s as unusual as or more unusual than the one observed. By inspection of the probabilities of occurrence listed in Table 31, we can see that samples

Table 31. DISTRIBUTION OF SAMPLE PROPORTIONS EXPECTED WHEN $p_u = q_u = .5$, AND $N = 6$

Number of A's	Proportion of A's (p_s)	Term of binomial	Proportion of samples	Probability of occurrence
0	.0000	q^6	$\frac{1}{64}$	.015625
1	.1667	$6q^5p$	$\frac{6}{64}$	.093750
2	.3333	$15q^4p^2$	$\frac{15}{64}$	.234375
3	.5000	$20q^3p^3$	$\frac{20}{64}$	.312500
4	.6667	$15q^2p^4$	$\frac{15}{64}$	.234375
5	.8333	$6q\,p^5$	$\frac{6}{64}$	.093750
6	1.0000	p^6	$\frac{1}{64}$	.015625

Source: Table 29.

with five or six males would be more unusual in the same direction as our observed sample and that a sample with two males would be as unusual in the other direction and samples with one or no males would be more unusual in the other direction. There is some dispute as to whether it is better to test a hypothesis by considering the probability of only the cases as unusual as or more unusual than the observed *in the same direction* as the observed deviation, or by considering the probability of *all* cases as unusual as or more unusual than the observed deviation *in both directions.* Either method is "right" if the results are carefully stated, but it is perhaps more common and appears to the writers to be more valid to

consider deviations in both directions. This is what we shall do unless there is a specific reason for doing otherwise.

Thus, to compute the probability required for our test, we add the proportions of samples with four males, five males, and six males, or we can add the probabilities, if we wish,

$$\frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{22}{64}$$

or $.234375 + .093750 + .015625 = .343750$

The probability $22/64$ or .343750 represents the fraction of sample proportions expected to deviate as far as or farther than our observed proportion in the *same* direction. We double this probability to get the probability of samples expected to deviate as far as or farther than our observed proportion on both sides of the universe proportion,

$$2\left(\frac{22}{64}\right) = \frac{44}{64} \text{ or } 2(.343750) = .687500$$

Then the object of this step is accomplished with the determination of .6875 as the probability that a sample proportion deviating as far as or farther than .6667 would be expected to be observed in the sampling distribution of proportions in samples of size six from a universe with a proportion of .5.[6]

4. *Nonrejection of hypothesis.* Since a sample proportion as unusual as the observed might be expected to occur seven times out of ten, we do not reject the null hypothesis. Note the limitations of a statistical test of a hypothesis: it does not tell us whether to reject or to confirm the hypothesis tested, it only tells us whether to reject or *not reject* the hypothesis. If the verdict is "not reject," or "accept" we still have no assurance that the hypothesis is correct. Certainly we have not proved from observing a sample proportion of .6667 that the universe proportion is equal to or less than .5. All we have done is to show that from a universe with a proportion of .5, we should expect to observe samples with proportions varying from the universe proportion as much as our observed sample proportion does in seven times out of ten.

5. *Interpretation of the test.* All we can conclude is that we do not know whether there are more males than females in the universe sampled. Our estimate of the population proportion is still .6, but we know that this is a very unreliable estimate and that its lower confidence limit

---

[6] If the sampling distribution described by the binominal expansion is not symmetrical about .5, that is, if $p \neq q$, the simplest way to obtain the desired probability is to add all the probabilities for different terms which are as small as the probability corresponding to the sample observed, or are smaller.

extends far below .5. We could test other hypotheses, such as $p_u \leq .4$ or $p_{u'} \leq .3$, etc. if we wished to investigate further, by evaluating the terms of the binomial with appropriate values of p and q as above. Perhaps the most important suggestion from the test just made is that if we wish to narrow the confidence interval around our estimate of the population proportion enough for us to tell whether or not there are more males than females in the universe, we must draw a larger sample.

Applications of the methods of this chapter to sociological problems. The problem of estimation of a universe proportion is met whenever there has been a sample study enumerating attributes. For instance, of the 1950 census questions which were asked of only the 20-percent sample, several can be regarded as the enumeration of attributes. In each case the proportion for the universe (the entire population of the United States) can be estimated to be the same as the proportion observed in the sample; and if (for purposes of simplification) the methods of drawing the sample are considered to be the equivalent of random methods, confidence limits for any degree of confidence desired can be set up to indicate the reliability or precision of the estimates.

The problem of testing hypotheses in the simplified form we have illustrated is not so commonly met in sociology as in some other fields. For we do not now have in sociology much theory highly enough developed to have deduced from it precise quantitative hypotheses about universe parameters to be tested in the way we have described. (As for instance geneticists test the appearance of traits in successive generations to see if they conform to the Mendelian theory of inheritance.) The much more frequent case in sociological work involves the testing of hypotheses relating to observations on two different samples, to establish the "significance" of a difference between them. Methods for making such tests will be given in Chapter 19, and while they are for different situations, they will be found to follow exactly the same steps outlined above. Since they are slightly more complex, however, it will be well for the student to master thoroughly the procedures presented in this chapter before going on to the tests which have more practical uses.

SUGGESTED READINGS

Snedecor, George W., *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), Chap. 1.

Yule, G. Udny, and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), Chap. 17.

Quantitative Distributions: Sampling Distributions of Measures of Central Tendency, Dispersion, and Form

Content of chapter. The distribution of an attribute (which units either possess or do not possess) in a universe is completely described when we know one parameter—the proportion of units in the universe possessing the attribute. The distribution of a quantitative variable (formed by the measures of a characteristic which the units may possess in varying degrees) in a universe requires a more complex description, usually including summarizing measures of its central tendency, its dispersion, and its form. In this chapter we shall present methods of forming estimates of the universe parameters describing the various aspects of a quantitative distribution, of stating the reliability of the estimates in several different ways, and of testing various hypotheses about the universe parameters. Then we shall consider the sampling distribution of the distribution as a whole, as manifested by the fluctuations in observed frequencies of class intervals, when samples of a given size are drawn from a distribution described by a frequency curve. To develop the methods, we shall again start with the case where the universe parameters are known, although our ultimate aim is to learn how to infer information about the universe from observations made on a sample of it, as is usually the situation in practical research problems.

SAMPLING DISTRIBUTIONS OF SUMMARIZING MEASURES OF CENTRAL TENDENCY

The sampling distribution of the arithmetic mean. Let us assume that the description of the distribution of a quantitative variable in a universe is known: that its form is normal, its mean is μ, and its standard deviation is σ. If we draw samples of size N from this universe and compute the mean, $\bar{X}$, for each sample, these means will vary from one

to another and from μ but will cluster around μ, as many above μ as below it. The distribution of these means is called the sampling distribution of means. This sampling distribution of means will also be of normal form with its mean at μ, but its standard deviation will be smaller than that of the original distribution. Specifically, the sampling distribution of means will have a standard deviation of $\dfrac{\sigma}{\sqrt{N}}$ where σ is the standard deviation of the original distribution in the universe and N is the number of cases in the sample. The standard deviation of the sampling distribution of means is called the standard error of the mean. It is given the symbol, $\sigma_{\bar{x}}$ or σ_{μ}, and defined, thus,

$$\sigma_{\bar{x}} \text{ or } \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{N}} \tag{1}$$

From inspection of formula (1) we see that the standard error of the mean, which is the primary measure of unreliability of an estimate of the universe mean, is larger when the original distribution has a greater degree of dispersion, and that it is smaller when the number of cases in the sample is greater. This is reasonable and confirmed by common sense. We know that an estimate of the mean of very similar units is more reliable than an estimate of the mean of very dissimilar units. We know also that we have more confidence in the reliability of an estimate based on a great number of cases than in an estimate based on very few cases. Thus this statistical concept of the standard error of the mean is simply a way of formulating more precisely the criteria for the reliability of an estimate which one already uses intuitively.

So far we have dealt with no approximations about the sampling distribution of means because if the original distribution is normal and if the universe parameters are known, the three items of information, that is, mean $= \mu$, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{N}}$, and form normal, describe exactly the distribution of the sample means. In actual practice, however, we usually do not know the mean, standard deviation, or form of the original distribution in the universe, and, therefore, our description of the sampling distribution of the means of samples is less exact. Or sometimes we know enough about the form of the parent or original distribution to know that it is not normal. We shall consider a practical example which illustrates the difficulties and then give the methods for treating them.

Example of a statistical inference about the universe mean. Let us assume that the 117 white tenant farm women on whom data were presented in Chapter 8 are a randomly drawn sample from all the white tenant farm women in the Tobacco Piedmont area of North Carolina.

(This assumption happens not to be correct, but for purposes of illustration, we can assume its correctness.) From the data on number of children ever borne by the 117 women we wish to estimate the mean number of children borne by all white tenant farm women in the North Carolina Tobacco Piedmont, who comprise the universe. We have no information on the distribution of number of children borne in the universe other than what we can infer from the sample.

Regardless of the form of the universe distribution, the best estimate of its mean is the observed mean of the sample, that is,

$$\hat{\mu} = \bar{X} \tag{2}$$

Evaluation of (2) with the value of the mean obtained on page 105 solves the first half of our problem of inference, that is, the estimation of the value of the universe parameter, thus,

$$\hat{\mu} = 6.34$$

The next part of the problem of inference is to investigate the unreliability of this estimate and for that purpose we wish to be able to describe the sampling distribution of means of samples of size 117. We cannot do this precisely because we do not know the universe parameters, but we shall use the following procedures for obtaining an approximate description.

To estimate the general nature of the form of the distribution in the universe, we can look at Figures 10 and 11 of Chapter 8, which show graphically the form of the distribution in the sample. It is an I-type curve, approaching normal but definitely skewed to the right, as the separation of the values of the mean and median indicate. Since this sample distribution includes the only information available about the universe distribution, we shall have to use it as an estimate of the universe distribution. Then we see that one of the conditions for the sampling distribution of means to be normal is not met—the orginal distribution is not normal. Fortunately, this is not a serious obstacle. While the original distribution must be normal for the sampling distribution to be *exactly* normal, the departure from normality in the sampling distribution will be much less than the departure from normal of the original distribution. We have mentioned the descriptive measure β_1 which can be used as a measure of skewness. Whatever the value of β_1 in the original distribution, its value in the sampling distribution of means will be only $\dfrac{1}{N}$ as much. In this case where N is 117, the skewness of the sampling distribution as measured by β_1 will be only $\dfrac{1}{117}$ or .008547 as much as the skewness of the original distribution and therefore is negligible. For while we did not

actually compute β_1 for this distribution, since its skewness was only moderate as shown by the coordinate chart, we know that its β_1 would be less than one and, therefore, that the β_1 of the sampling distribution of means of samples of 117 would be less than .008547. In general, if N is not very small, we can assume that the distribution of sample means is near enough to normality for all practical purposes if the original distribution approximates a normal distribution only very roughly—that is, if it has only one mode somewhere near the center and if the ends taper off to zero.

We do not know the mean of the sampling distribution of means, for μ is what we are trying to estimate; and if we used our estimate, $\hat{\mu} = \bar{X}$ in testing the reliability of the estimate, we should be proceeding in a circular manner and would arrive nowhere. But if we knew the standard deviation of the sampling distribution, $\sigma_{\bar{x}}$ (the standard error of the mean), we would find it useful to describe the dispersion of the sampling distribution, whatever its mean is. To get $\sigma_{\bar{x}}$, however, we need to know σ, the standard deviation of the original distribution in the universe, and that too is unknown. We shall have to estimate its value from information gained from the sample and use this estimate in investigating the reliability of the estimate of the mean.

While $\bar{X}$ is the best estimate of μ, s is not usually regarded as the best estimate of σ, since it is biased in the direction of being too small. It is generally agreed [1] that it is better to use the following estimate,

$$\hat{\sigma} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N-1}} = \sqrt{\frac{\Sigma x^2}{N-1}} \tag{3}$$

This expression differs from that for s,

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{\Sigma x^2}{N}} \tag{4}$$

only in having $N - 1$ instead of N in the denominator. If s has already been computed, $\hat{\sigma}$ can be computed from it by the relation,

$$\hat{\sigma} = s\sqrt{\frac{N}{N-1}} \tag{5}$$

Now we wanted an estimate of σ to substitute in the expression,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \tag{1}$$

When we use an estimate of σ instead of the actual universe parameter, we shall have to indicate by placing a circumflex over $\sigma_{\bar{x}}$ that we have obtained only an estimate of it, thus,

[1] See pp. 260–261.

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{N}} \tag{6}$$

By substituting the right member of (3) in (6), we get

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\Sigma x^2}{N(N-1)}} \tag{7}$$

By substituting the right member of (5) in (6), we get

$$\hat{\sigma}_{\bar{x}} = \frac{s\sqrt{\dfrac{N}{N-1}}}{\sqrt{N}} = \frac{s}{\sqrt{N-1}} \tag{8}$$

We may use either (6), (7), or (8) for computing purposes, as they give identical results. Since authors vary so greatly in notation, the student should differentiate clearly between the definitions of σ, s, and $\hat{\sigma}$ as given here, in order to avoid confusion over what might appear to be contradictory formulas in other texts.

Using data from page 123 in Chapter 9, we substitute in (5) to get an estimate of the universe standard deviation,

$$\hat{\sigma} = s\sqrt{\frac{N}{N-1}} = 3.395\sqrt{\frac{117}{116}} = 3.4096$$

Substituting this value of $\hat{\sigma}$ in (6) to get an estimate of the standard error of the mean, we have

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{3.4096}{\sqrt{117}} = .3152$$

There is an important difference between the sampling distribution of a mean (a summarizing measure of a quantitative distribution) and the sampling distribution of a proportion (a summarizing measure of a nonquantitative distribution). The standard deviation of the sampling distribution of a proportion is different for each different universe value of a proportion, whereas the standard deviation of the sampling distribution of a mean is independent of the universe value of μ, although dependent on the universe value of σ. We can, then, set up confidence limits for the estimate of the mean by the first procedure given in Chapter 15 without involving the approximation made in the case of setting up confidence limits for the estimate of a proportion. There is an approximation here also, but it is of a different nature. It is the use of the estimate of the universe value of σ rather than the actual value itself. We shall first assume that it is permissible to use $\hat{\sigma}$ for σ and later give a more accurate procedure.

The process of determining the 95-percent confidence limits of the

estimate of the universe mean consists of finding two values, M_1 and M_2, one above $\hat{\mu}$ and one below $\hat{\mu}$ at distances of 1.96 standard deviations from the estimate.[2] If the universe value, μ, is outside the confidence limits, the probability that we would observe an $\bar{X}$ as unusual as the one we did is less than .05. The values are computed as follows:

$$M_1 = \bar{X} - 1.96\ \hat{\sigma}_{\bar{x}}$$
$$M_1 = 6.34 - 1.96(.3152) = 6.34 - .6178 = 5.722$$
$$M_2 = \bar{X} + 1.96\ \hat{\sigma}_{\bar{x}}$$
$$M_2 = 6.34 + 1.96(.3152) = 6.34 + .6178 = 6.958$$

The graphic interpretation of these limits is shown in Figure 35. If the universe mean is anywhere between 5.72 and 6.96, the probability of observing a sample mean as unusual as $\bar{X}$ is greater than .05. If the universe mean is outside of the two limiting positions shown in Figure 35,

Number of children borne
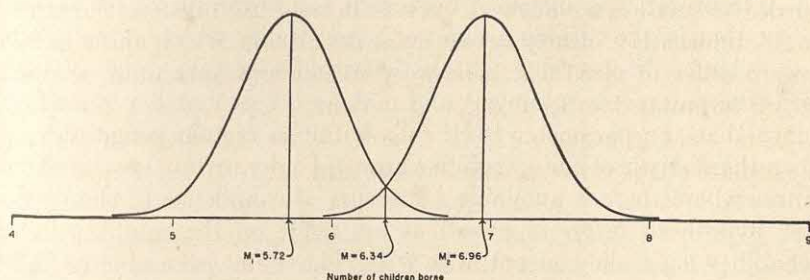
Figure 35. Estimate of the Universe Mean, $\hat{\mu} = \bar{X} = 6.34$, with 95-percent Confidence Limits, $M_1 = 5.72$, $M_2 = 6.96$. M_1 and M_2 are the limiting values of the means of universes from which the probability of observing a mean as unusual as 6.34 in a sample of 117 is equal to or greater than .05. (Source: pp. 250–251.)

the probability of observing a sample mean as unusual as the observed $\bar{X}$ is less than .05. As before, the correct interpretation of confidence limits is that if we were to continue to draw samples and determine confidence limits in this way, our confidence limits would include the universe mean in 95 out of 100 times. This interpretation is subject to a qualification, however, since we have been forced to use an estimate rather than the actual value of the universe standard deviation and hence have obtained only estimates of the confidence limits.

Incorrect interpretations of reliability of estimates. At the risk of confusing the student, we feel it is wise to point out the reasons why the interpretations of confidence intervals (or ranges similar to them) found in older texts and still prevalent in much research reporting are incorrect. These incorrect interpretations are usually given along with a range of

---

[2] Note that we cannot use the more logical notation μ_1 and μ_2 because of the confusion that would arise from the similarity to the notation for moments.

±2 or ±3 times the standard error of the mean. The ranges are obtained in a manner similar to that we have just described for determining the 95- or 99-percent confidence range, except that different multiples of the standard error of the mean are used—or more precisely, different multiples of *estimates* of the standard error of the mean are used. The incorrect interpretations are in general of two sorts, and we shall illustrate both of them with the same data used to illustrate confidence limits and then show why they are wrong.

The first sort of incorrect interpretation is as follows.

The mean of our sample is 6.34, and the standard error of the mean is .315. Therefore, the probability is .95 that the universe mean lies within the range 6.34 ± 2$\sigma_{\bar{x}}$ or from 5.71 to 6.97.

This interpretation is incorrect because it is an example of *inverse probability*, that is, the turning around of a probability which actually refers to the chance of observing a statistic within a certain range of values (when the parameter is known) and making the probability refer to the chance that the parameter itself falls within a certain range of values (when the statistic of one sample is known). Under certain special circumstances where there is available a measure of confidence in the truth of some hypothesis *before* observations are made on the samples, inverse probability is a valid concept. The writers have not been able to find an example in sociological research of a situation in which inverse probability is applicable. Therefore, the above type of interpretation is not recommended in the research problems now dealt with in sociological research.

The second sort of incorrect interpretation is as follows.

The mean of our sample is 6.34, and the standard error of the mean is .315. Therefore, in successive random samples of the same size from this same universe we should expect 95 out of 100 means to fall within the range 6.34 ± 2$\sigma_{\bar{x}}$ or from 5.71 to 6.97.

A slight variation of the last sentence may be often seen as,

Therefore, the probability is .95 that the mean of a second random sample of the same size from this same universe will be between 5.71 and 6.97.

Or slightly differently,

The probability that the mean of a second random sample of the same size from this same universe will be less than 5.71 or greater than 6.97 is .05.

All three variations of this incorrect interpretation are wrong for the same reason. They are all based on the assumption that our estimate of the universe mean, $\hat{\mu} = \bar{X}$, coincides *exactly* with the universe mean, μ, that is, on the assumption that from this one sample we have determined the value of the universe parameter with absolutely no margin of error.

This assumption is contradictory to all the theory and practice regarding sampling, for we do *not* expect our estimates to be *exactly* correct; that is why we are trying to measure their unreliability. If, however, this assumption were permissible, the three variations of interpretation given above are satisfactory. For all of them state probabilities of observing values of statistics of samples when the universe parameter is known.

Since this sort of incorrect interpretation is so prevalent in research reporting, let us examine the *direction* of the error involved in it. When from one sample statistic we are trying to estimate the value of the corresponding statistic in another sample, sampling error enters twice into the process—once in the difference between the observed statistic and the universe parameter, and again in the difference between the universe parameter and the second sample statistic. The above incorrect interpretation takes account of only the second of these differences and therefore errs in the direction of *under*stating the difference to be expected. The interpretation, then, exaggerates the reliability of estimates, and, consequently, because of its direction of error is an even more serious mistake.

The above incorrect interpretations are applied not only in the case of means but also in the cases of many other statistics—correlation coefficients, regression coefficients, and especially to differences between means and correlation coefficients. The same unacceptable assumption of having made a perfect estimate underlies the interpretation in all these cases. With the correct procedures and interpretations now available from developments in modern statistics, the use of such incorrect interpretations is no longer excusable in any of the cases.

Student's distribution. The procedure we have explained for setting up confidence limits assumes that sample means are normally distributed with a standard deviation of $\hat{\sigma}_{\bar{x}}$. Actually, when $\hat{\sigma}$ is used instead of σ and consequently $\hat{\sigma}_{\bar{x}}$ is obtained rather than $\sigma_{\bar{x}}$, the sampling distribution of the ratio of the observed mean to its estimated standard error is not exactly normal. If $N > 30$, the departure is not important, but if $N < 30$, the departure becomes important enough to require the use of a special table giving the areas under a curve which differs from the normal in form. Since the form of the distribution is slightly different for each different value of N, values are tabled for each value of N, or rather for each value of

$$n = N - 1 \qquad (9)$$

where n represents the number of degrees of freedom, a concept which will be treated later. The group of distributions is called "Student's Distribution" because they were published under the pseudonym of "Student" in 1908 by W. S. Gossett, who, because of his position with a

commercial company, did not publish under his own name. Appendix Table D gives these tables for certain selected values of P (the probability that a deviation so unusual would be observed) and for values of n from 1 to 30. For instance, if our above mean were computed from a sample of 17 instead of 117 and if we wish to find the multiple of the estimated standard error of the mean, beyond which only 5 percent of the cases would fall, we look up the entry corresponding to $P = .05$ and $n = N - 1 = 17 - 1 = 16$. We find this value is 2.120 as compared with 1.96 for the normal distribution. These t values, as the values from Student's distribution are called, are always larger for a required level of probability than corresponding values in the normal distribution, although as N becomes infinitely large, the t values approach the normal values. The student should not be confused over the fact that the table looks so different from the table of areas under the normal curve, because it is arranged differently. The table of areas under the normal curve has the multiples of standard deviation units in the argument, and the half areas (probability that a deviation *not* more unusual in the same direction would be observed) as entries in the cells; while the t table has two sets of arguments, n (degrees of freedom) and P (probability that a deviation as unusual as or more unusual than would be observed) with the multiples of standard deviation units as entries in the cells. Certain tables of areas under the normal curve are arranged this way also.

The use of the t distribution instead of the normal distribution affords what is called an "exact" test of a hypothesis about the mean, for the t distribution describes *exactly* the sampling distribution of $\dfrac{\bar{X} - \mu}{\hat{\sigma}_{\bar{X}}}$, just as the normal distribution describes exactly the sampling distribution of $\dfrac{\bar{X} - \mu}{\sigma_{\bar{X}}}$. This point needs clarification because we have heretofore been dealing with sampling distributions of single statistics, not of a ratio of two statistics. One should note first that there are two ways of interpreting the fact that the normal distribution describes the sampling distribution of the mean. One way is that it describes the sampling distribution of the *mean itself* when it is expressed as a deviation from the universe parameter in units of $\sigma_{\bar{X}}$. Another way is that it describes the sampling distribution of the *ratio* to the deviation of the observed mean from the universe mean to the standard error of the mean. This ratio has a standard error of one, and, therefore, we can look up its corresponding probability by entering Appendix Table A with the actual value of the ratio. In fact, Table A is sometimes called a table of areas corresponding to a normal deviate with unit variance.

These different interpretations make no difference when the standard error of the mean $\sigma_{\bar{X}}$ is known. The second interpretation, however,

clarifies somewhat the situation when an estimate of $\sigma_{\bar{x}}$ has to be used. When in the denominator of the ratio we substitute not $\sigma_{\bar{x}}$ but an estimate of it, the variance (and standard deviation) of the sampling distribution becomes greater than one and the form of the distribution is different. The amount of change in the size of the standard deviation and in the form of the distribution is different for each different size of sample (for each different number of degrees of freedom). Therefore, we cannot draw one frequency curve representing the t distribution as we can for the normal distribution because the t distribution is different for each N. As N becomes large, the corresponding t distribution approaches normality. Although the use of the t distribution is always more correct theoretically than the use of the normal distribution in setting up confidence limits or in testing hypotheses about the mean when the standard error of the mean must be estimated, there is no practical reason for using it unless $N < 100$, and little difference in results unless $N < 30$. To illustrate the small amount of difference when $N = 117$ as in our example, we shall compare the confidence limits secured by using the t distribution and those secured by using the normal distribution. We shall have to show the values of the confidence limits to more than two decimal places, however, for when rounded to two decimal places, the two pairs of confidence limits are the same. From the table of Student's distribution we find by interpolation that the multiple corresponding to $P = .05$ and $n = 116$ is 1.98 instead of 1.96 as in the normal distribution. The confidence limits are computed thus,

$$M_1 = \bar{X} - 1.98(.3152) = 6.34 - .62410 = 5.716$$
$$M_2 = \bar{X} + 1.98(.3152) = 6.34 + .62410 = 6.964$$

The confidence limits based on the multiples obtained from the t distribution determine a confidence interval which is just slightly broader than that determined by the confidence limits secured by use of the normal distribution, which we now show to four decimal places,

$$M_1 = 5.722$$
$$M_2 = 6.958$$

Although the difference between the two pairs of confidence limits is negligible in this case, let us note that the use of the normal distribution always gives a confidence interval which is too narrow, and that when N is small, it is very important to use the t distribution so that we do not overestimate the reliability of our estimates.

Application of the above procedures in sociological research. The estimation of universe means from samples is a procedure of wide applicability in sociological research. For example, the sample used in the Current Population Survey of the Bureau of the Census consists of about

25,000 households. From information obtained from these households estimates are made of various characteristics of the population of the entire United States. Approximate 95-percent confidence limits are estimated by appropriate adaptations of the methods presented. Means of samples have long been used as estimates of universe values, but it is only recently that the custom of publishing confidence limits or confidence ranges with such estimates has been introduced.

Test of a hypothesis about the universe mean. As in the case of universe proportions, sociological theory has not developed to the stage where many specific quantitative hypotheses concerning universe means can be deduced to be tested by the results of observations from samples. Instead, the hypotheses tested are more usually limiting cases, such as $\mu = 0$, or are arrived at empirically from observation of another sample. The second case is the more frequent, but its treatment will be delayed to Chapter 19, where hypotheses concerning the distributions of the same characteristic in two groups will be tested.

For an illustration, let us consider a situation which is a sort of hybrid between the two cases of theoretical hypotheses about the universe and simple comparison of two samples. Let us suppose that from the 1940 census sample question, "number of children ever borne," the results were tabulated separately for the Tobacco Piedmont Subregion of North Carolina, and within this area by residence, race, and tenure. Let us further suppose that from the sample so obtained of white tenant farm women of the area, an estimate was made of the number of children ever borne for the whole group. Since there are about 17,000 women in the specified universe, and since the sample includes 5 percent, the estimate would be based upon approximately 850 observations. If the standard deviation were no larger than that in our example, we should expect the estimate of the mean number of children ever borne in the universe to be more reliable than in the case of the 117 women of our illustration, since the number of cases is more than seven times as great. (Since the standard error of the mean varies inversely with the square root of the number of cases, the standard error of the mean estimated by the census sample should be roughly $\sqrt{\dfrac{1}{7}}$ or about one third the size of that for our sample, if the standard deviations of the distributions are not greatly different.) Furthermore, since the Bureau of the Census tried to develop a sampling technique which would give results that would approximate those that would have been obtained through random sampling, there are other reasons for having more confidence in the estimate made from the census data.

If, then, we take the hypothetical census estimate as the universe parameter, we can use this description of the universe as a hypothesis

and test it to see if our sample mean would be likely to be observed in a sample of 117 from such a universe. Since we *know* the sample did come from such a universe, if the test leads to rejection of the hypothesis, we shall interpret it to mean that the sample was not drawn in such a way as to be representative of the universe. Thus, we shall test a hypothesis about the mean of the universe to throw light on the representativeness of the sample obtained by our sampling techniques.

Data on 5-percent sample questions of the census are not available by counties, and therefore we cannot actually make this test. But let us assume that the census has released as an estimate of the mean number of children ever borne by white tenant farm women in the area the figure 5.44, with an estimate of the standard deviation of the distribution as 3.0. Let us proceed to the problem we have suggested by the steps outlined in Chapter 15 for testing a hypothesis.

1. *Formulation of the hypothesis.* This time we shall not assume that our sample was randomly drawn but shall include the method of drawing the sample in the hypothesis. This is what usually happens in practical problems—our hypothesis tested is complex, and we cannot always tell which feature of the hypothesis leads to rejection. The hypothesis to be tested includes two major parts: (a) that the observed sample was drawn from the universe described in such a way as to be an acceptable approximation to random methods; (b) that the description of the universe is as follows: $\mu = 5.44$, $\sigma = 3.0$, and form roughly approximating normal.

2. *Description of the sampling distribution expected of means in samples of 117.* Since we know both the mean and the standard deviation of the universe, and since we know the distribution roughly approximates the normal, we shall expect the mean of the sampling distribution of means of samples of 117 to be equal to the mean of the universe, 5.44. We shall expect the standard deviation of the sampling distribution to be

$$\sigma_{\bar{X}} = \frac{3.0}{\sqrt{117}} = .28$$

We shall expect the form of the sampling distribution to be a closer approximation to the normal than is the parent universe, close enough that the tables based on the normal curve may be used.

3. *Determination of the probability that a sample mean as unusual as 6.34 would be observed in this sampling distribution.* We first express the deviation of the observed mean from the universe mean in standard deviation units of the sampling distribution of means,

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{6.34 - 5.44}{.28} = \frac{.9}{.28} = 3.2 \text{ standard deviation units}$$

In Appendix Table C we find the entry corresponding to 3.2 is .00137. This means that the probability of observing in a sample of 117 a mean as unusual as (or more unusual than) 6.34 is .00137.

4. *Rejection of hypothesis.* Since a mean as unusual as our observed mean would be expected to occur less than two out of a thousand times, we reject the hypothesis that our sample was randomly (or approximately randomly) drawn from a universe with a mean of 5.44 and a standard deviation of 3.0.

5. *Interpretation of the test.* By the rules of the game the hypothesis has to be rejected because P is so small. It must be remembered, however, that the hypothesis involves two parts, the description of the method of drawing the sample and the description of the universe. The rejection of the hypothesis may be necessitated by the incorrectness of either or both of these parts. Usually we approximate random methods of drawing a sample as best we can and assume the truth of the first part in order to test the second part. In this particular example just the reverse is the case. We know that the second part is true. That is, every one of the 117 women was actually a member of the defined universe (and we have assumed this universe to be described by the estimates of the larger census sample.) Therefore, the rejection of the hypothesis must be interpreted to mean that the method of selecting the sample was such as to give results less representative in number of children ever borne than a purely random method would have done. We can say that the sample is biased in the direction of including too many women with more than the average number of children. Although the supposed census estimates are purely hypothetical data, such was actually the case. With one exception only women with children were selected for the sample, and there was an undue proportion of those with large families, since the study was focused on women of high fertility.

This illustration introduces some of the complexities involved in testing statistical hypotheses. Almost always our results have to be qualified in such a way as to state that a decision on one point is possible only if we assume certain things on other points. The practical methods of selecting samples are implicitly assumed to yield samples as representative as a purely random method in almost every test regarding values of universe parameters. It is clear, then, that we cannot explore thoroughly the subject of interpretation of tests of hypotheses until after these approximations have been considered in the next chapter.

The sampling distribution of other measures of central tendency. The median, the mode, the geometric mean, and the harmonic mean, as well as the arithmetic mean, all have sampling distributions. We shall not be concerned with the sampling distributions of the geometric and harmonic means, however, since they are used very rarely as measures of central

tendency in sociological research. Nor shall we consider the sampling distributions of the median and mode so fully as we have the sampling distribution of the arithmetic mean, both because they are less important measures of central tendency and because their sampling distributions are more difficult to describe.

The sampling distribution of the median has as its mean the universe value of the median. But the standard deviation of the sampling distribution (the standard error of the median) varies in size according to the form of the original distribution of measures. Whereas the standard error of the mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \qquad (1)$$

not only for strictly normal distributions, but also for distributions departing markedly from normal, the standard error of the median varies with the kurtosis of the distribution from which the sample is drawn. For a normal distribution it is

$$\sigma_{Md} = 1.25331 \frac{\sigma}{\sqrt{N}} \qquad (10)$$

If the original distribution is platykurtic, σ_{Md} is larger than is indicated by (10), and if the original distribution is leptokurtic, σ_{Md} is smaller than indicated by (10). Thus, in general, for distributions approaching normality the standard error of the median is greater than that of the mean, although this is not true for an extremely leptokurtic distribution.

For a given distribution and given size of sample if one of two statistics, which are summarizing measures of the same aspect of the distribution, has a smaller standard error than the other, it is said to be the more *stable* statistic. Thus, generally, the mean is a more stable summarizing measure than the median. Stability is, of course, a desirable feature for a statistic because when the statistic is used as an estimate of the corresponding universe parameter, the confidence interval around the estimate will be narrower for a more stable statistic. The greater stability of the mean as compared with the median or the mode is one of its advantages which causes us to use it more frequently than the other two as a summarizing measure of central tendency of a distribution.

Not only is stability a desirable feature of a statistic, but the measurability of the degree of instability of a statistic is also of considerable importance. For the mean the standard error is $\frac{\sigma}{\sqrt{N}}$ to a close approximation even though the form of the original distribution is quite different from normal. For the median the standard error is dependent upon the form of the original distribution, and it is only when we know the sum-

marizing measures of that form and when they happen to have certain values that we can obtain the value of the standard error of the median. Therefore, it is often not possible to test precisely hypotheses about the median.

The mode has the same disadvantages as the median in a greater degree. Generally, it is even more unstable than the mean or median, and its standard error is also dependent on the form of the original distribution and cannot ordinarily be determined precisely. In fact, it is only when the mode is determined by finding the point on the X scale where the height of a fitted curve is the maximum that any precise description can be made of its sampling distribution. Such methods are beyond the range of this text, and we shall have to leave the mode with the warning that it is a relatively unstable measure even when computed precisely and that it is even more unstable when computed by the approximate procedures given in Chapter 8.

SAMPLING DISTRIBUTIONS OF SUMMARIZING MEASURES OF DISPERSION

The sampling distribution of the standard deviation. We have referred to the fact that the standard deviation of the sample is not the best estimate of the standard deviation of the universe. We should qualify this statement by adding that the criteria for "best" estimate are not completely agreed upon and that the choice of criteria determines whether s or $\hat{\sigma}$ as defined above is the "best" estimate. A discussion of the controversial matter, particularly of the different points of view of R. A. Fisher and J. Neyman, may be found in one of J. Neyman's lectures, "Statistical Estimation."[3] All writers are agreed, however, that s is a biased estimate of σ. This can be demonstrated without any elaborate mathematics. An unbiased estimate is one which approaches the universe parameter as a limit when the number of samples is increased indefinitely. The standard deviation of any group of measures in a sample is less than their root mean square deviation from any point which is not their sample mean. Then if we should attempt to estimate the universe standard deviation from the mean of the sampling distribution of a number of s's from many samples, our estimate will be too small because the sum of the squared deviations of observed measures from their sample means will be less than the sum of their deviations from the universe mean. The amount of bias can be determined by mathematical derivation, and the estimate can be corrected for this bias by using $N - 1$ rather than N in the denominator of $\hat{\sigma}$. It is evident that when N is greater than 100, the difference between N and

---

[3] *Lectures and Conferences on Mathematical Statistics* (Washington: Graduate School of the United States Department of Agriculture, 1938), pp. 127–142.

$N - 1$ is so small that the correction becomes unimportant and may be omitted, although it is always theoretically correct to use it.

If the original or parent distribution of the characteristic is normal and if N is large, the sampling distribution of $\acute{\sigma}$ approximates a normal distribution with a mean equal to σ and a standard deviation,

$$\sigma_{\acute{\sigma}} = \frac{\sigma}{\sqrt{2\,N}} \tag{11}$$

If instead of using the universe σ in this expression, we have to substitute $\acute{\sigma}$, the formula for the estimate of the standard error of the standard deviation

$$\acute{\sigma}_{\theta} = \frac{\acute{\sigma}}{\sqrt{2\,N}} \tag{12}$$

Or in terms of s, this becomes,

$$\acute{\sigma}_{\theta} = \frac{s}{\sqrt{2(N-1)}} \tag{13}$$

If we know the parent distribution is normal and if $N > 100$, we can set up confidence limits or test hypotheses by using $\acute{\sigma}_{\theta}$ as the standard deviation of the sampling distribution which will have an approximately normal form. However, any departure from normality of the parent distribution affects the size of the standard deviation of the sampling distribution and its form more seriously than in the case of the sampling distribution of the mean. Even if the parent distribution is perfectly symmetrical, but platykurtic or leptokurtic, formulas (11), (12), (13) do not give accurate values for the standard error of the standard deviation. And if the number of cases is small, the sampling distribution is positively skewed. Therefore, the table of normal areas is not frequently used with estimates or hypotheses relating to the standard deviation. Instead the more accurate description of the form of the distribution with its appropriate table is used, that is, the *chi square* distribution, which, like the t distribution, is dependent upon $n = N - 1$. The *chi square* distribution will be discussed in the next section of this chapter. Furthermore, especially since the more recent advances in statistics inaugurated chiefly by R. A. Fisher, more interest attaches to the sampling distribution of the variance than of the standard deviation. Because in practice we are usually interested in hypotheses concerning two or more estimates of variance, the special methods for handling the ratio of estimates of variance to be presented in Part IV in the chapter on analysis of variance are of more value in testing hypotheses than the methods dealing with single estimates. Therefore, the treatment of the matter of reliability of estimates of σ will be left somewhat incomplete in this section.

SAMPLING DISTRIBUTIONS OF SUMMARIZING
MEASURES OF FORM

In Chapter 9 we used only graphic presentation to test the normality of our distribution, but frequently we wish to make a more precise test of the normality of a distribution. Such a test can be made using measures derived from β_1 and β_2 computed in Chapter 14. These measures are gamma coefficients and are defined as follows:

$$\gamma_1 = \sqrt{\beta_1} \tag{14}$$

$$\gamma_2 = \beta_2 - 3 \tag{15}$$

These two gamma coefficients have sampling distributions which can be derived from a consideration of the sampling distributions of the moments on which they are based. As we have already noted in the case of the standard deviation, which is based on the second moment, the sampling distribution of measures based on the higher moments are different for different forms of the original distribution. The expressions for the standard errors of the gamma coefficients reduce to a relatively simple form only when the form of the original distribution is normal. In this particular case the standard error of γ_1 is

$$\sigma_{\gamma_1} = \sqrt{\frac{6}{N}} \tag{16}$$

and the standard error of γ_2 is

$$\sigma_{\gamma_2} = \sqrt{\frac{24}{N}} \tag{17}$$

In a normally distributed universe where $\gamma_1 = 0$ and $\gamma_2 = 0$, the means of the two sampling distributions are both zero, and the forms of both sampling distributions are normal if N is large. Therefore, we can test the hypotheses that the universe from which our set of observations may be considered a random sample is symmetrical ($\gamma_1 = 0$) and that it is mesokurtic ($\gamma_2 = 0$). As an example we shall make these tests for the distribution shown in Table 26. Computing the values of the gamma coefficients we have

$$\gamma_1 = \sqrt{1.3224} = 1.15$$

$$\gamma_2 = 5.867 - 3 = 2.867$$

The standard error of γ_1 is dependent only on the number of cases, 99, and we find it by evaluating formula (16), thus,

$$\sigma_{\gamma_1} = \sqrt{\frac{6}{99}} = .246$$

The deviation of the observed γ_1 expressed in units of the standard deviation of the sampling distribution is

$$\frac{1.15 - 0}{.246} = 4.674$$

While this exact value is not tabulated in Appendix Table C, from the tabulated value of the probability of 4.5 in Appendix Table C we can say that the probability of getting a sample with a γ_1 as unusual as this from a normal parent distribution is less than .000007. This probability is so small that we reject the hypothesis that the universe is symmetrical and without testing γ_2 can assert that this is not a random sample from a normally distributed universe.

We will test the hypothesis that $\gamma_2 = 0$, however, in order to illustrate the method. Substituting the formula (17) we find that

$$\sigma_{\gamma_2} = \sqrt{\frac{24}{99}} = .49$$

The deviation of the observed γ_2 expressed in units of the standard deviation of the sampling distribution is

$$\frac{2.867 - 0}{.49} = 5.85$$

Referring this value to Appendix Table C, we find that the probability of obtaining a γ_2 as unusual as 2.867 in random samples of 99 from a normal distribution is less than .0000006. This is additional evidence that this is not a random sample from a normally distributed universe.

All summarizing measures of form, whether β_1 and β_2 or functions of the betas such as γ_1 and γ_2, are based on the higher moments. The higher the moment, the less "stable" it is. Therefore, the standard errors of the betas and gammas are great unless N is large. Tests of departure from normality such as those just illustrated are not very meaningful if N is less than 100, and the precision of statistical inferences about the form of the universe distribution is quite limited unless N is much greater than 100.

SAMPLING DISTRIBUTION OF THE FREQUENCIES OF A GROUPED QUANTITATIVE DISTRIBUTION

Frequency distributions. When N measures have been observed with regard to the degree of incidence of a characteristic manifested by N varying units, the series of measures is called a quantitative distribution. We have learned to describe such quantitative distributions by computing

summarizing measures of central tendency, dispersion, and form. We have learned also to combine these summarizing measures into one algebraic equation for describing the quantitative distribution as a whole, although we have considered in this text only the case where the form can be assumed to be normal. Before we did this, however, we learned by grouping measures to form a frequency distribution (which gave a description of the quantitative distribution as a whole) acceptable for presentation purposes although less precise than an algebraic equation.

The case where the universe distribution is assumed to be known. In investigating the sampling distribution of frequency distributions it is again necessary to consider first the case where the universe parameters are known. In the preceding chapter we said that the quantitative characteristic "percentage of males in a sample of 100" is normally distributed. Note that "maleness" is a nonquantitative characteristic, but that the "percentage of males in a sample of 100" is a quantitative characteristic. The infinite universe of all possible samples of size 100 has this characteristic distributed according to the following description,

$$\mu = 60.0$$
$$\sigma = 4.899$$
$$\gamma_1 = 0$$
$$\gamma_2 = 0$$

Or if we wish to describe algebraically the frequency distribution we should expect in groups of 100 samples grouped into class intervals of two units, we can substitute the above summarizing measures into equation (3) of Chapter 14 and obtain the following equation.

$$Y_c = 16.2833e^{\frac{-(X-60.0)}{2(4.899)^2}} \qquad (18)$$

By use of the appendix tables it is possible to find the exact number of observations we should expect in each class interval of percentage males. (The computation of expected frequencies will be explained shortly.) The series of expected frequencies corresponds to the mean of an ordinary sampling distribution. We know that if we drew successive groups of 100 samples from this universe, we should not "expect" the series of frequencies in each group to be identical with the "expected" frequencies because of sampling fluctuation. The problem which faces us, then, is that of finding one summarizing measure of the discrepancy between the series of observed and expected frequencies and, further, of describing the sampling distribution of this summarizing measure.

Chi square as a measure of a frequency distribution's departure from expectation. The measure we shall use which combines information on the deviations of all the observed frequencies from their corresponding

expected frequencies is called chi square, χ^2. The distribution of chi square is known, and on the basis of the distribution Appendix Table E has been derived. Like the t table the chi square table reveals what the respective probabilities are that values as unusual as those obtained would be observed under specified hypotheses. If the probability is large, we say that we "accept" the hypothesis of the description of the universe; if the probability is small, we "reject" the hypothesis of the description of the universe.

Test of a hypothesis describing completely the universe distribution. The hypothesis tested in such a case is a more complex hypothesis than those we have been testing above, which bear a relation to the values of only one or two universe parameters. The more complex hypothesis is a complete description of the distribution in the universe, which may be stated in terms of the four parameters μ, σ, γ_1 and γ_2, or in terms of the equation (18). Under the specifications of the hypothesis as to mean, standard deviation, and form of a distribution we compute the frequencies expected in the size of sample and class intervals that we have used. Then, we compute chi square, which measures the combined deviations of the series of frequencies from expectations. Finally, we determine from Appendix Table E what the probability is that a chi square as unusual as that evaluated would be observed and on the basis of the probability either accept or reject the hypothesis.

If we find reason to accept the hypothesis, the interpretation is that the observed frequency distribution *may* have come from the universe described. If we find reason to reject the hypothesis, the interpretation is that the evidence is against our believing that the observed frequency distribution came from the universe described. Since the description of the universe involves several aspects, we may attempt to determine which part of the description is responsible for rejection of the composite description.

In assigning the blame to a particular aspect it is necessary to follow certain fundamentals underlying the fitting of the normal curve as explained in Chapter 14. It can be demonstrated theoretically that the *normal* equation which has its mean, standard deviation, and number of cases equivalent to those of an observed distribution is the *normal* distribution for which the test outlined above will give the smallest chi square, or the greatest P. Therefore, if we find reason to reject the description of this particular normal distribution, there will be even greater reason for rejecting all other normal distributions. Thus, we can blame the aspect of *form* for rejection, when rejection is the verdict.

Because the aspect of form is the part of the composite description on which the test as we have outlined it gives information, the test has long been known as the chi square test of "goodness of fit" of an estimated

equation and its corresponding curve to the observed frequency distribution and to its corresponding graphic presentation. We shall test the goodness of fit of only a normal curve because this test is the simplest and is the one most frequently needed, although chi square is not limited to testing the fit of this particular form. In fact, the chi square distribution is useful in many different sorts of tests of hypotheses, a number of which will be presented in the succeeding chapters of this book.

We shall now present the actual procedures for testing the hypothesis that the universe from which the frequency distribution of Table 28 may be considered a random sample of 100 observations is a normal distribution ($\gamma_1 = 0$, $\gamma_2 = 0$) with a mean of 60.0 and a standard deviation of 4.899. (Note that we are changing from proportions to percentages, and that the measurement of each sample of size 100 is called an observation, and the entire group of such observations is called a sample.) The test may be interpreted graphically as testing the fit of the superimposed normal curve in Figure 33 to the coordinate chart of the observed distribution. We shall present the test in the customary five steps. Although the content of these steps may seem somewhat different from that of previous tests, the statistical function of each step is the same.

1. *Formulation of the hypothesis to be tested.* The most concise statement of the hypothesis to be tested is that the incidence of males in the universe from which the grouped data of Table 28 may be considered a random sample of 100 observations is described by the equation,

$$Y_c = 16.2833e^{\frac{-(X - 60.0)}{2(4.899)^2}}$$

From this equation we must find the actual frequencies expected in the 2-unit class intervals used in Table 28. This procedure is known as computing the expected frequencies. We include it in the step of formulating the hypothesis because it is really an extension of the process of deduction whereby we determine what results the formulated hypothesis would lead us to expect.

We do not actually substitute in equation (18) but make use of the relations between the abscissa and the area of the normal curve, which are tabulated in Appendix Table A. The first six columns of Table 32 show the computations for obtaining the expected class interval frequencies. Instructions for obtaining the entries in each column are found at the head of the column. The entries in column (5) are the differences between successive entries in column (4), except that the entry for the interval containing the mean is the sum of the two entries in column (4). The entries in column (6), obtained by multiplying the entries in column (5) by 100, are the "expected" frequencies, which have been deduced from the

description of the universe distribution for a particular size of sample and set of class intervals.

2. *Description of the sampling distribution of the frequency distribution.* Since our interest is in the sampling distribution, not of one summarizing measure, but of a set of frequencies, its description is more complicated than in previous tests. In fact, we shall have to devise some one summarizing measure which can be used to show the extent of the deviations of frequencies in samples from the corresponding expected frequencies. The measure developed for this purpose is the sum of a series of quantities which are similar to the percentage deviation of each frequency. In fact, they are the sum of the proportions of deviations with each proportion weighted by the value of the deviation. If for a class interval,

$$f_c = \text{expected (or computed) frequency}$$
$$f = \text{observed frequency in a sample of } N;$$

then the quantities referred to can be represented by the expressions

$$\frac{(f - f_c)}{f_c}(f - f_c) = \frac{(f - f_c)^2}{f_c}$$

The sum of these quantities is the desired summarizing measure of deviation, chi square,

$$\chi^2 = \sum \frac{(f - f_c)^2}{f_c} \tag{19}$$

If correspondence between expected and observed frequencies were perfect, each $f - f_c$ would be equal to zero and therefore χ^2 would be zero; but such a coincidence would be rare indeed. Appendix Table E describes the sampling distribution of χ^2, by showing the value of chi square corresponding to selected values of P.

3. *Determination of the probability that a set of deviations in this sampling distribution would yield a chi square as unusual as that obtained.* First, we must compute the chi square for our sample set of frequencies. Intervals at the extremes have been combined so that no f_c will be smaller than 5. Columns (7) to (10) in Table 32 contain the necessary computations in terms of the symbols which have just been explained. The sum of the entries in column (10) is chi square, 2.8962. Before we can refer this value to Appendix Table E, we must take into account the number of degrees of freedom on which it is based. Furthermore, it must be kept in mind that in order to have frequencies for comparison which are not too small, some of the class intervals have been combined, resulting finally in nine class intervals, as shown in column (8). (Class intervals should usually be combined until no expected frequency is less than 5.) Our chi square test is thus based on a comparison of nine pairs of frequencies.

Table 32. COMPUTATIONS FOR TESTING THE HYPOTHESIS THAT THE UNI-
VERSE FROM WHICH THE DATA OF TABLE 28 ARE A RANDOM SAMPLE IS
DESCRIBED BY THE NORMAL EQUATION $Y_c = 16.2833e^{\frac{-(X-60.0)}{2(4.899)^2}}$

True limits		Deviation of class limits $(1) - \bar{X}$ (2)	Deviations of class limits in σ units $(2) \div \sigma$ (3)	Proportion of area between c.l. and $\bar{X}$ Appendix Table A (4)	Proportion of area lying between successive class limits (5)
Lower c.l. (1)	Upper				
$-\infty$		$-\infty$	$-\infty$	.50000	.00295
46.5		-13.5	-2.7556	.49705	.00650
48.5		-11.5	-2.3474	.49055	.01681
50.5		-9.5	-1.9391	.47374	.03663
52.5		-7.5	-1.5309	.43711	.06771
54.5		-5.5	-1.1226	.36940	.10388
56.5		-3.5	-0.7144	.26552	.14534
58.5		-1.5	-0.3061	.12018 ⎱	.16078
	60.5	0.5	0.1020	.04060 ⎰	
	62.5	2.5	.05103	.19501	.15441
	64.5	4.5	0.9185	.32081	.12580
	66.5	6.5	1.3268	.40775	.08694
	68.5	8.5	1.7350	.45865	.05090
	70.5	10.5	2.1432	.48393	.02528
	72.5	12.5	2.5515	.49466	.01073
	∞	∞	∞	.50000	.00534
Sums					1.00000

Source: Table 28.

The number of degrees of freedom involved in any test of a hypothesis is equivalent to the number of *independent* observations on which the test is based. To compute the degrees of freedom in this situation we begin with nine, the number of comparisons to which the test relates. By fitting a normal curve to observed data we have imposed certain relations between the various expected frequencies and the observed frequencies which keep the comparisons from being independent. For each condition we impose on the expected distribution we sacrifice one degree of freedom if the restricting condition is determined by the observed distribution. The one condition we have imposed in this situation is that the sum of the expected frequencies equal the sum of the observed frequencies. This costs us one degree of freedom.

We have not made the mean and standard deviation of our fitted normal curve coincide with the mean and standard deviation of our observed distribution because we got these parameters from other considerations.

Table 32. COMPUTATIONS FOR TESTING THE HYPOTHESIS THAT THE UNIVERSE FROM WHICH THE DATA OF TABLE 28 ARE A RANDOM SAMPLE IS DESCRIBED BY THE NORMAL EQUATION $Y_c = 16.2833 e^{\frac{-(X-60.0)}{2(4.899)^2}}$

Number of observations in samples of 100 expected to fall within class intervals $f_c = 100 \times (5)$ (6)	Observed frequency f (7)	$f - f_c$ (8)	$(f - f_c)^2$ (9)	$\dfrac{(f - f_c)^2}{f_c}$ (10)
.295 ⎫ .650 ⎪ 1.681 ⎬ 6.289 1.681 3.663 ⎭	0 ⎫ 1 ⎪ 2 ⎬ 6 3 ⎭	−0.289	.08352	.0133
6.771	6	−0.771	.59444	.0878
10.388	7	−3.388	11.47854	1.1050
14.534	15	0.466	.21716	.0149
16.078	20	3.922	15.38208	.9567
15.441	16	0.559	.31248	.0202
12.580	12	−0.580	.33640	.0267
8.694	7	−1.694	2.86964	.3301
5.090 ⎫ 2.528 ⎪ 1.073 ⎬ 9.225 .534 ⎭	7 ⎫ 3 ⎪ 1 ⎬ 11 0 ⎭	1.775	3.15062	.3415
100.000	100			2.8962

Had we fitted our normal curve using these summarizing measures from our observed distribution, as is frequently done, this would have cost us two more degrees of freedom (and would have yielded a smaller chi square value). The requirements that both γ_1 and γ_2 be zero were derived from theoretical considerations and not from the observations, so imposing these requirements does not sacrifice any degrees of freedom.

In this situation, then, we have sacrificed only one degree of freedom, leaving our chi square of 2.8962 to be based on eight degrees of freedom. In Appendix Table E we find that for 8 degrees of freedom ($n = 8$) the value of chi square corresponding to a P of .95 is 2.733, and the value of chi square corresponding to a P of .90 is 3.490. Hence, we know that the P corresponding to chi square of 2.8962 will lie between .90 and .95. Interpolation is not necessary and we can write our results thus,

$$.90 < P < .95$$

Or in words we can say that the probability that a chi square as unusual

as 2.8962 would be observed under the hypothesis is between .90 and .95.[4]

4. *Nonrejection of hypothesis.* Since the probability is far greater than any of the customary levels of significance, we cannot reject the hypothesis.

5. *Interpretation of test.* The observed distribution may have come from a normal distribution with a mean and standard deviation as specified. Theory tells us that the parent universe from which this particular distribution of samples was drawn is a normal universe with mean and standard deviation as specified. Although our particular group of samples did not have the identical mean or standard deviation expected in the parent universe, the distribution of the sample values is such that we cannot reject our hypothesis regarding the distribution of the universe. Therefore, we may say that the normal curve shown in Figure 33 fits the distribution satisfactorily in interpreting graphically our test as a test of goodness of fit. We emphasize again, however, that one must keep in mind the logic involved and remember that we have not proved that the sample of 100 observations is drawn from a normally distributed universe, *nor have we confirmed the hypothesis that the universe is normal.* We have merely shown that frequency distributions departing from expectation as far as our observed distribution or further are common occurrences in random samples drawn from a normal distribution.

The applications of tests of goodness of fit in sociological research. In a thorough analysis and description of a quantitative distribution, the aspect of form must be treated just as the aspects of central tendency and dispersion are treated, and if there is to be a generalization of the results, tests of hypothesis about form must be made. Hypotheses about β_1 and β_2 (or about γ_1 and γ_2) can be tested separately, or the test of normality of form just explained can be made.

Although logically these tests are as important as those relating to the universe values of the means of distribution, they have not yet become practically important in sociological research problems. Years ago Franklin Henry Giddings suggested the interpretation that skewness of a distribution might mean that purposeful plans of man had been superimposed upon "natural law" which would in itself produce a normal distribution.[5] Almost all characteristics associated with income for individuals or with plane of living for demographic areas show a marked skewness to the right. Although the tests above explain, among other things, the process of determining whether such observed skewness is

---

[4] Whenever one gets a probability greater than .95, the computations, sampling procedure, etc., should be carefully checked. Probabilities greater than .95 are as unusual as probabilities less than .05 and frequently indicate an error in computations or procedure.

[5] Franklin Henry Giddings, *The Scientific Study of Human Society* (Chapel Hill: University of North Carolina Press, 1924), pp. 142–143.

great enough to warrant the conclusion that the observed sample did not come from a normally distributed universe, they leave many of the questions regarding form unanswered. The field of investigation into the aspect of form of the distributions of sociological characteristics has been greatly neglected.

One reason for this neglect is that the techniques for investigating form are simple and easy to apply only if the form is normal or near normal. Since many of the distributions of sociological interest vary so markedly from normal, these simplified techniques cannot be employed. Let us examine why so many of the distributions are not normal.

For a distribution to approach normality closely the range of its possible values must extend several standard deviation units on either side of the mean. For many characteristics, the measures of which can take only positive values, the range is cut off on the left side within two or three standard deviation units of the mean, causing a skew to the right. This is shown graphically in the distribution of children borne in Figure 13. For characteristics distributed among demographic areas the location of zero on the scale of the measure, or the location of some other limiting value (as in the case of death rates), often produces a skew, usually to the right. This results in a departure from normality, of course, often to such a degree that some of the methods of analysis of relationship between two distributions are not applicable, since they are based upon the assumption of normality. This matter will be treated more fully in Chapter 20.

SUGGESTED READINGS

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), pp. 305–317.

Dixon, Wilfrid J., and Massey, Jr., Frank J., *Introduction to Statistical Analysis* (New York: McGraw-Hill, 1951) Chap. 9.

Lindquist, E. F., *Statistical Analysis in Educational Research* (Boston: Houghton, 1940), Chaps. 2 and 3.

Peatman, John Gray, *Descriptive and Sampling Statistics* (New York: Harper, 1947), Chap. 13.

Sampling in Social Research: General Principles and Methods, Problems of Interpretation

Relation of chapters of Part III. We are proceeding with the presentation of inductive statistics along two main lines: the one, explanation of the specific procedures whereby we make estimates, compute measures of their unreliability, and test hypotheses; the other, explanation of the general principles underlying these procedures, their applicability in social research, and interpretation of their application. Neither aspect can be completely treated before the other is considered, and, therefore, the chapters of this part alternate in direction of emphasis. The chapter on statistical induction introduced the more general problem and principles; the next three chapters were occupied with specific procedures; this chapter will go back to general considerations underlying sampling and its interpretation is social research; the next chapter will treat more specifically the application of sampling in social surveys; and the remaining chapters will turn to specific statistical procedures for testing hypotheses in social research.

OBJECTIVE OF SAMPLING

The general methods of sampling have been developed to achieve the objective of selecting from a universe of varying units a group of units in such a way that generalizations can be drawn from observations on the sample to the universe from which it was drawn with a measurable probability of error. Three criteria are relevant for achieving the objective: (1) the sample must represent the universe (that is, it must be unbiased); (2) the sample must be of adequate size to produce reliable results (that is, as measured in terms of a specified range of error); (3) the sample should be designed in such a way as to be efficient (that is, in comparison with alternate designs that could have been used).

As indicated earlier, problems of application and interpretation of

sampling theory differ in some respects according to the general type of situation in which they are met. One may be sampling from an existent, finite universe, seeking to draw inferences from the sample about that universe. Or, one may be sampling from a hypothetical universe, with or without sampling from an existent, finite universe. In the immediately following sections of this chapter, the actual methods of sampling by which the above listed criteria are met (or attempted) will be presented in terms of the situation of sampling from a finite, existent universe. In a later section on interpretation of sampling the case of sampling from a hypothetical, infinite universe will be treated.

SIMPLE SAMPLING

The term "simple sampling" is used for the situation in which a sample is to be drawn from a universe without any prior subdivision of the universe into strata (groups of relatively similar units). We wish to consider methods that will achieve the objective described above or that will approximate the objective. The concept of a random sample is basic to an understanding of modern sampling practice.

Random sampling. A "random" sample is one drawn from a universe in such a way that every unit in the universe has an equal probability of being drawn into the sample and that the inclusion of any one unit does not affect the probability of any other unit's being drawn into the sample. In a random sample we say that chance alone determines which of the units in the universe enter into the sample. The factor of "chance" is not easily defined in practical terms. Chance is generally regarded as the result of a multiplicity of factors simultaneously operating independently to produce "random" variation. We try to make models to illustrate the operation of chance by such means as drawing cards which are thoroughly shuffled between each draw, by throwing dice which are carefully shaken between each throw, or by other ways where the intention of the drawer cannot influence the choice of units either consciously or unconsciously, and where the conditions of equal probability of drawing any unit and of independence of the various draws are approximated.

At best these methods of drawing "random" samples are only approximate models of the mathematically defined concepts on the basis of which the theory of probability has been developed. We wish to emphasize this approximate nature at the outset, for in sociological research we get even farther away by approximating these approximations, a fact which one must bear in mind when making interpretations of the results obtained from sampling.

Sampling by means of Tippett's Random Sampling Numbers. A method which yields very successful approximations to the theoretical

concept of drawing a sample randomly involves the use of a list of about 40,000 four-digit numbers which were made from census digits thoroughly scrambled or shuffled by L. H. C. Tippett.[1] Since the use of this method has been demonstrated by empirical research to give excellent results, it is generally recommended for any situation where it is applicable, and it will, therefore, be described in some detail. The situation required for its application is that there be available a list of all the sampling units in a universe. The method for drawing a "random" sample of any given size from the list of units consists of the following steps:

1. Assign numbers serially to the units in the list; it does not matter in what order they are numbered, whether alphabetically, geographically, or in order of their measures on some characteristic.

2. From the number of units in the universe, determine whether it will be necessary to use one-, two-, three-, four-, or more digit numbers. For instance, if there are 5,000 units in the universe, one should use four-digit numbers, if 500 three-digit numbers, and so on.

3. By a toss of a coin, or other procedure not governed by the intention of the person drawing the sample, determine the place in the list of Tippett's numbers for starting and then read off successive numbers of the digit length required as they occur in some arbitrarily selected direction—up, down, left, right, or diagonally.

4. For each of Tippett's numbers read off, draw for inclusion in the sample the unit in the listed universe corresponding to that number. If there is no corresponding unit, ignore the number; if a number is repeated, ignore its second occurrence. Continue this process until a sample of the required size is drawn.

It is urged that this procedure for drawing samples from finite universes be followed wherever lists of sampling units are available when a simple random sample is required. Where there are no such lists, other methods for approximating a random sample will have to be used.

Reasons for random sampling. Now that one practical method for approximating a random drawing of a sample has been explained, let us make clear the reasons why a simple random sample is desired. Is it because a simple random sample always affords the closest replica of the universe in the several aspects of distribution of the characteristic in which we are interested? No. Given certain previous knowledge, one may employ other methods of drawing a sample which will yield a more perfect miniature of the parent universe. Then why do we so often wish to draw simple *random* samples? First, because we often do not have the previous knowledge that would enable us to draw a sample that would yield

---

[1] *Tracts for Computers*, No. 15, Random Sampling Numbers, arranged by L. H. C. Tippett (London: Cambridge University Press, 1927).

estimates about the universe with a probability of smaller error than a simple random sample, and without such knowledge the random method is our best bet for securing representativeness. Second, even when we have such knowledge and on the basis of it divide our universe into parts or strata from which certain numbers of units of the sample are to be drawn, it is best to use random selection within these parts or strata for securing representativeness within them. And finally, we shall be able to use the theory of probability in interpreting our results only if we can expect sampling error of the degree which occurs in random sampling (or some specified modification or adaptation thereof for which allowance can be made). Only if chance is allowed to determine which units are included in a simple sample, is the expected distribution of the error in estimates derived from the sample charted for us. And even though approximation after approximation to ideal conditions of random sampling is used, it has been found profitable to try to approach random methods nearly enough to utilize the expectations of results derived by the theory of probability. Only with such methods can we obtain information of the precision of our estimates or can we test hypotheses about the universe.

Practical difficulties in obtaining random samples. Facing the fact squarely that we shall have to be content with approximations to random samples, let us examine some of the situations actually met in sociological research. First, there is the case where a list of sampling units is available and the procedure described above for using Tippett's random sampling numbers is applicable. For instance, suppose we wish to draw a random sample of counties in the United States. We should simply number them, probably alphabetically under states with the states arranged either alphabetically or geographically, and proceed as described above. Or suppose we wish to draw a sample of all telephone subscribers in Chicago. We should assign numbers serially to the names of persons or firms listed in the telephone directory and proceed similarly. Since, however, we do not have a continuous register of persons or families for various subareas of the United States, this procedure is very often not feasible for the type of research sociologists wish to do. A common mistake in such a case is to use a list of only part of the universe when a list of the whole universe is not available. For instance, suppose we wish to sample all the families in Chicago and proceed to draw our sample as described from the list of residence telephone subscribers. It is immediately evident that we will get an overrepresentation of the higher economic class families and an underrepresentation of the lower economic class families. Then for any characteristic associated with economic level, our sample will not be representative. When approximations such as this one cause the sample to depart from representativeness, we say that they introduce a bias. The

chief criterion for the acceptability of approximations in methods is that they do not introduce biases, or if they do, that we estimate the biases and correct for them.

Selection at regular intervals.   Where there are no lists of sampling units available, the ingenuity of the research person is demanded to construct some approximate method. A much used alternative to the numbering of the sampling units is the procedure of selecting units for the sample at regular intervals when they have some definite spatial arrangement. For instance, in the problem of sampling city families one might plan to follow a certain route covering every street in the city and to select every tenth, twentieth, or fiftieth household. In residential areas of detached one-family houses, this procedure would be rather simple to apply but in areas of multiple dwelling unit structures an additional step of first listing the dwelling units in each structure would be required. In other situations, the characteristic being studied might be associated with spatial intervals, and a bias might be introduced.[2] Adapting this procedure to rural areas involves further practical difficulties. Rural dwellings are usually not in orderly rows and many of them are out of sight of public roads so that following an orderly scheme of selection is often difficult. Although this method, or some modification of it, was formerly used fairly frequently by rural sociologists, the development of "area sampling" to be described later has largely replaced the method of selection at regular intervals in surveys of rural families.

Hit-or-miss sampling.   Because of the confusion over the meaning of the term "random," it may be thought that unplanned, haphazard, hit-or-miss methods might yield the equivalent of a random sample. It is true that they *might*, but it is also true that especially in the sort of sampling a sociologist does, they usually *do not*. Hit-or-miss methods are likely to be affected by several sorts of bias. First, there is the matter of *teleological bias;* a person with an axe to grind may be influenced either consciously or unconsciously to choose the units which would prove his case; or if he is very conscientious, he may overcorrect to the extent of not selecting the units which would prove his case. Secondly, there is the matter of *dispersion bias;* a person trying honestly to select a random sample may show a tendency to notice and select more of the extreme cases than those in the middle range; or knowing of the prevalence of this tendency, he may overcorrect and select an undue proportion of cases near the average, in either case obtaining a bias in dispersion. Thirdly, there is the matter of *accessibility bias;* a person is likely to select for his sample those units which are easiest to get at, causing a bias if there is an association between the accessibility of the sampling units (especially where they are spatially

---

[2] See G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), p. 371.

distributed) and the distribution of the characteristic studied. Now these three sorts of bias may operate in either direction according to the person doing the selecting of the sample and the nature of the sampling situation. And while a person trained to avoid these biases in particular situations might be able to draw a fairly representative sample (see comments on public opinion polls below), in general the results from hit-or-miss methods are unpredictable and, therefore, are not acceptable as a source of information for scientific research. We wish to emphasize that we have no assurance that unplanned, hit-or-miss methods of sampling will give results equivalent of those of carefully designed random sampling, even when the sampler has the purest motives and thinks he is controlling every possible source of bias. Unfortunately, even in research reporting, the term "random" is sometimes incorrectly used to describe such hit-or-miss methods.

STRATIFIED SAMPLING

The term "stratified sampling" does not mean that the methods of simple random sampling are not involved. On the contrary, it usually means the selecting of a group of random samples, one from each of several strata or parts of the universe. The term "stratified sampling" is used in contradistinction to the term "simple sampling" in which one sample is drawn from the universe as a whole. We referred earlier to the fact that if we are given certain previous knowledge, we may be able to draw a sample that will provide estimates for the universe that have smaller sampling errors than corresponding estimates from a simple random sample. One way of utilizing such knowledge results in the process of stratified sampling. Because by stratified sampling we get more reliable estimates of universe parameters from the same number of observations, it is a more efficient method of sampling than the simple random methods and, if feasible, should be used when the necessary previous knowledge is available.

The group of subsamples comprising a stratified sample may each have the same number of units or they may have differing numbers of units. When the numbers of units in the subsamples are proportional to the numbers of units in the strata, the method of sampling is called "proportionate stratified sampling," and this method has been widely used. For greatest efficiency, however, the variation of units within the strata should be taken into consideration in determining the number of units to be drawn from each stratum.

Let us consider an oversimplified hypothetical case. Suppose we wish to draw a sample of children from a certain school in order to study the distribution of the characteristic age. Imagine for a moment that from

school records the age distribution of the children in the school is known. Now if we divide the required number of cases in the sample proportionately to the frequency of children in 1-year age groups and select the indicated number randomly from each 1-year age group, we shall get a sample much more representative as to age of the universe of school children than we should expect to get by simple random sampling. Obviously, this case is absurd because if we knew the age distribution of the universe, we should not be interested in drawing a sample for the purpose of estimating it. Suppose, however, that we do not know the age distribution of the children but that we do know the grade distribution, a reasonable enough supposition. Now age and grade are said to be associated or correlated characteristics because an older child is likely to be in a higher grade. If we stratify our universe by grades and choose randomly from each grade a number of children proportional to the number in that grade, we shall again secure a proportionate stratified sample that provides an estimate of the age distribution of all the school children with a smaller sampling error than we should expect from a simple random sample. Whenever we know the distribution of a characteristic associated with the one we are interested in studying, we can utilize this knowledge to advantage by stratifying the universe on the basis of the known distribution and drawing from the several strata random samples proportional to the size of the strata. The composite sample formed from the several random samples is called a proportionate stratified sample. Each stratum may consist of all the units in a certain class interval of a quantitative characteristic, or it may consist of all the units in a certain category of a nonquantitative characteristic.

When for purposes of sampling, a universe is divided into strata which are the categories or class intervals of an associated characteristic, the associated characteristic is called a "control." The more closely associated the control characteristic is with the one being studied, the greater the efficiency of the sample design using this control is over a simple random sample. It is possible to use more than one characteristic for control in stratification by methods referred to later.

For computing estimates and measures of reliability from the stratified sample, which is a composite of several random samples, certain modifications in the formulas and procedures of Chapters 14 and 15 are needed. All of the formulas and procedures, however, are based upon the assumption that the samples within strata are *randomly* drawn. Again, any departure from randomness in method of selecting a sample within a stratum is an approximation and must be critically examined. In actual practice departures are common. Also common in research practice and reporting are the applications of methods of estimation assuming randomness in situations where there were only gross approximations to

randomness, with no qualifications expressed in the interpretation of the results.

More complex sample designs. Since the middle of the 1930's there have been important developments in sampling design, particularly for large-scale surveys. Examples of these will be mentioned in the next chapter. Some of the more important additional methods used include: (1) two or more levels of sampling, such as selecting a sample of counties and then selecting a sample of areas or households within the sample counties; (2) cluster or small-area sampling, in which the smallest unit used in sampling contains several of the units used in analysis; (3) design of sample so that results can be tabulated by classes for which independent estimates exist, with a resulting reduction in sampling error; (4) combination of random and systematic (every *n*th unit) methods to obtain geographic dispersion, especially in rural areas.

Stratified nonrandom sampling. It is obvious that if the association between the control characteristic and the investigated characteristic is very close, the strata which consist of categories or class intervals of the control characteristic will have very little variation in the investigated characteristic. When a stratum is relatively homogeneous with respect to the investigated characteristic, it does not make much difference which of its units are drawn for a sample. When several controls, each closely associated with the investigated characteristic, are used, the strata are even more likely to be homogeneous with respect to the investigated characteristic. It then follows naturally that when the strata are very homogeneous, the selection of units within the strata becomes less important as a source of bias, but the amount of bias introduced is generally unknown.

This is the general principle of "quota sampling" upon which public opinion polls and many commercial surveys have largely operated. Following the failure of the polls to forecast the results of the 1948 election, a good deal of attention was given to the method of sampling used, and a critical evaluation is listed in the readings at the end of this chapter. We shall not attempt to present such methods in detail here, for it seems preferable to devote the space available to those methods that can be recommended with fewer qualifications.

SIZE OF SAMPLES

How large should a sample be? If there is one question asked a consulting statistician more frequently than ony other, it is, "how large a sample should I take?" Often one hears such glib answers as, "Never less than 100," or "500," or "At least five percent." However, the question is actually unanswerable until the following items of information are

given: the designation of the parameters which one wishes to estimate, the range of unreliability permissible in estimates, and a rough estimate of the dispersion of the investigated characteristic. If these are known, it is a relatively simple matter to compute approximately the number of cases required for a simple random sample.

Example of determination of number of cases required in a sample. Suppose we wish to estimate the incomes of farmers within an area with a 95-percent confidence range of $400. We may have no very precise estimate of the standard deviation of the distribution of farm income in the area, but we may know that the range is from zero to about $6,000. For fairly large samples with approximately normal distributions, the standard deviation can be approximately estimated as one fourth to one sixth of the range, which would here give an estimate of about $1,250. Now in setting up the 95-percent confidence limits of the estimate of the universe mean, we should plan to use the relations,

$$\bar{X} - M_1 = 1.96 \hat{\sigma}_{\bar{x}} \tag{1}$$
$$M_2 - \bar{X} = 1.96 \hat{\sigma}_{\bar{x}}$$

If the confidence range is to be $400, the confidence limits M_1 and M_2 must be at a distance of $200 on either side of the sample mean $\bar{X}$. Substituting this value in the first equation, we have,

$$\$200 = 1.96 \hat{\sigma}_{\bar{x}} \tag{2}$$

Now we must estimate the standard error of the mean from our estimate of the standard deviation of the universe thus,

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{\$1,250}{\sqrt{N}} \tag{3}$$

Substituting (3) in (2) we have an equation with only N unknown,

$$\$200 = 1.96 \times \frac{\$1,250}{\sqrt{N}} \tag{4}$$

Solving (4) for N, we have,

$$\sqrt{N} = \frac{1.96 \times 1,250}{200} = 13$$
$$N = 170 \text{ (approximately)}$$

Thus we see that 170 is an estimate of the size of sample required to estimate the mean farm income with a 95-percent confidence range of $400 if the standard deviation of the distribution of income in the area is approximately $1,250, and if the distribution of farm incomes is normal.

This is a very rough estimate based upon rough approximations, one of which (that income is normally distributed) we know is far from true. The estimate should not be taken too literally because of the incorrectness of the approximations on which it is based, and often it is well to increase the number in the sample above that found to compensate for the approximations. The above procedures, however, do give us a usable method of anticipating our results, and they can be invaluable in planning research projects. If in this investigation, one wishes a narrower confidence range or a higher level of confidence, a larger number of cases will be necessary, as can be found by appropriate substitutions. Or if the range in income were twice as great, $12,000 instead of $6,000, the estimate of the standard deviation of the universe would be twice as great, and to counterbalance it, N would have to be 2^2 or 4 times as great, or 680 to estimate a mean with a 95-percent confidence interval of only $400.

A similar procedure can be used to determine the number of cases needed for estimation of a proportion of units possessing a certain attribute, within a given confidence range. In this case an estimate of the proportion in the universe is needed in order to estimate the standard error of the proportion. If we have no estimate of this, however, we can compute an estimate of the standard error of a proportion using p_u as .5, for this is the value which gives a maximum standard error. Then we can be sure that the N we determine will be at least large enough to meet the requirements we have specified.

For example, we may wish to know what size sample we need to draw in a public opinion poll in order to have a 95-percent confidence range of only a 3-percent vacillation on either side of our estimate of the percentage of voters favoring a certain political candidate. If we use the value of the universe proportion favoring candidate A, p_u, as .5, we shall obtain the maximum number needed. Then from the relation,

$$\sigma_{p_u} = \sqrt{\frac{p_u q_u}{N}} \tag{5}$$

and the relation,

$$p_1 - {}_{p_u} = 1.96 \hat{\sigma}_{p_u} = .03 \tag{6}$$

we can obtain N. First we solve (6) for $\sigma_{\hat{p}}$,

$$\sigma_{\hat{p}_u} = \frac{.03}{1.96} = .0153$$

Next we substitute $\sigma_{\hat{p}_u} = .015$ and $p_u = q_u = .5$ in (5),

$$.015 = \sqrt{\frac{(.5)(.5)}{N}}$$

and solve for N,

$$.015 \sqrt{N} = .5$$

$$\sqrt{N} = \frac{.5}{.015} = 33 +$$

$$N = 1,000 \text{ (approximately)}$$

Thus, we see that a sample of approximately 1,000 will meet our requirements. In the technical supplement to an article by S. S. Wilks, there is a chart for determining approximately the number of cases required for different values of the universe proportion and of the 99-percent confidence range.[3]

Such methods, with adaptations and extensions to take into account the design of the sample, are the only way of answering correctly the question of how large a sample should be. The answer always depends on what degree of reliability is required for the purpose of the research and on the amount of dispersion in the distribution of the characteristic studied if it is quantitative, or on the proportion of incidence if it is nonquantitative. If the universe is very homogeneous with respect to a certain quantitative characteristic, a quite small sample may yield more reliable results in the estimation of the parameters describing this distribution than a much larger sample of another universe which is very heterogeneous with respect to the characteristic studied. Therefore, rule-of-thumb answers of arbitrary numbers or percentages are misleading, and one should insist on being given the information required, even if in the form of extremely rough estimates, before attempting to answer the question.

The problem of determining the size of a sample becomes more complex when the sample is to be stratified, but the general principles of determining the size of the sample are the same as those illustrated for simple random sampling.

A difficult complication in the problem of determining the size of a sample is met in many practical situations where not a single characteristic but many characteristics are being investigated simultaneously. If expense is not the governing consideration, the size of sample may be determined by the number required to yield the precision desired for estimates relating to the distribution of the most important characteristic, or the greatest number required by the same considerations applied to several important characteristics or to all characteristics. Note that we stated above the qualification, "if expense is not the governing consideration." In actual research situations expense usually is the governing consideration and the major determinant of the number of cases studied in

[3] S. S. Wilks, "Confidence Limits and Critical Differences between Percentages," *The Public Opinion Quarterly*, 4 (June 1940), p. 333.

samples. However, even when this is true, there is usually a certain amount of flexibility which permits the exercise of judgment based on computations similar to the above. For instance, although the financial outlay for field work may already be determined, one may have the choice of whether the money should be spent collecting data on numerous characteristics from fewer units or on a smaller number of characteristics from more units. In such a case decisions as to the degree of unreliability of estimates permissible should be made first for the several distributions and then the number of cases required computed as above. A complex problem of this type involving double sampling has been treated by J. Neyman.[4]

Reliability in small samples. Along with the development of small sample theory, which makes possible fairly precise estimates of the unreliability of estimates made from small samples, has been a corresponding development of the application of this theory to actual research situations. Small samples are being utilized more frequently, especially in stratified sampling with several controls where a high degree of homogeneity is secured within the various strata. There is still prevalent among many research workers, however, a blanket distrust of any results based upon small samples. This distrust has a reasonable basis, of course, to the extent that reliability is always increased by using a larger sample. Number of cases, however, is not the only determinant of reliability, and the critics of small sampling theory and practice often overlook this point. They also often overlook the fact that the degree of reliability required varies with the purposes of a research project and that in certain situations a wide confidence interval may be permissible. At any rate, those who use small samples have just as valid a claim to correct scientific procedure as those who use large samples if they compute the measures of unreliability of their estimates and present them along with their results. They may still find themselves criticized simply on the basis of "too few cases" by those who really do not understand all the factors entering into the question of the validity of a sample, but they will also find an increasing number of modern statisticians ready to defend them.

It is true that certain modifications have to be made in some of the procedures of securing estimates of unreliability and testing hypotheses if the sample is small. Certain of the tables of probability are derived from approximations which are valid only when N is large enough for us to assume that certain discontinuous distributions approach normality, or for us to neglect terms involving N^2 or higher powers of N in the denominator. However, for small samples, inaccuracies in the approximations are lessened by the use of Student's distribution instead of the normal in

[4] J. Neyman, "Contribution to the Theory of Sampling Human Populations," *Journal of the American Statistical Association*, 33 (March 1938), pp. 101–116.

certain situations, the use of a binomial expansion instead of the normal distribution in others, and the restriction of chi square tables to situations where the number of cases expected is not less than 5 in others. Therefore, we urge that sweeping criticisms made solely on the basis of smallness of numbers in a sample be carefully examined before being taken too seriously.

SAMPLING FROM A FINITE UNIVERSE

Both the theory and the procedures of sampling we have presented so far have been developed for the situation where the universe is theoretically unlimited or infinite. The property of an infinite universe on which such theory and procedures depend is that when one unit has been selected for the sample and withdrawn from the universe, its removal does not affect the compositon of the universe as to distribution of the investigated characteristic. Obviously, this is true only if the number of units in each category or class interval is infinite. A finite universe, while not actually having this property, approximates it if the number of units in a sample is very small compared with the total number of units in the universe. For in such a case, even after all the sample units are drawn, the effect on the distribution of the characteristic in the universe is inappreciable. In the models constructed to illustrate sampling, as for instance urns with white and black balls in them, the effect on the composition of the finite universe of removing a unit when it is drawn for a sample is eliminated by replacing the drawn unit after data on it have been recorded. Such a procedure is known as "sampling with replacements," and it results in making sampling from a limited universe conform with the expectations of sampling from an unlimited universe, since the proportional frequencies of categories or class intervals in the universe remain the same throughout the sampling process.

In sociological research situations sampling with replacements is rarely the procedure used. Yet, in the situation of practical sampling from an existent universe, which we are discussing now, the universe is always limited. Therefore, the theory developed for sampling from infinite universes is never exactly appropriate to the practical sociological situations of sampling from limited universes without replacements. If the expression,

$$\sqrt{\frac{P - N}{P}} \tag{7}$$

where P = number of units in the universe
and N = number of units in the sample,
is not appreciably different from one in value, then the theory and pro-

cedures developed for sampling from infinite universes may be applied to sampling from the finite universe. If N is less than 20 percent of P, except in very refined work, the formulas and procedures for infinite universes are used without correction.[5] For instance, when N is $.1P$, the value for expression (7) becomes

$$\sqrt{\frac{P - .1P}{P}} = \sqrt{\frac{.9P}{P}} = \sqrt{.9} = .949$$

which is close to one. The reason that the nearness to the value one of expression (7) is taken as a criterion for using the methods developed for infinite universes is that this expression enters as a correction factor in the formula for the estimate of the standard deviation of the sampling distribution of estimates for the mean and in formulas for standard errors of other statistics. Thus, for a finite universe, the standard error of the mean is

$$\sigma_{\bar{x}_f} = \sigma_{\bar{x}} \sqrt{\frac{P - N}{P}} \tag{8}$$

where P = number of units in universe
$\quad N$ = number of units in sample
$\quad f$ = subscript denoting that the measure is for a finite universe.

Since $\sqrt{\dfrac{P - N}{P}} < 1$

we see that the measures of error for estimates based on a sample from a finite universe are always smaller where the magnitude of N is appreciable compared with that for P than they would be for estimates based on a sample of the same size from an infinite universe. Therefore, one usually errs on the side of conservatism if he uses the formulas developed for an infinite universe in such a situation. (Yet see discussion of errors of the first and second kinds in Chapter 19.) In other words, when our sample comprises an appreciable part of the limited universe, the unreliability of estimates is less than in the case of a sample of the same size from an unlimited universe, and if we use the unlimited universe formulas, we exaggerate the measures of unreliability of our estimates. But the exaggeration is not important unless the sample is one tenth or one fifth or more of the universe, and in most practical situations it does not need to be corrected. Moreover, even when we are sampling from an existent, finite universe, it is often true that we wish to make estimates and test hypotheses relating to the infinite universe of possibliities, to be discussed in the next section; and in this case, we should use the formulas for

[5] T. J. Woofter, Jr., "Common Errors in Sampling," *Social Forces*, 11 (May 1933), pp. 521–525.

infinite universes even though the sample comprises more than one tenth of the existent universe.

SAMPLING FROM A HYPOTHETICAL UNIVERSE

Observation of all units in a limited universe. Let us approach the concept of a hypothetical universe by means of an example. Suppose we are interested in investigating some demographic characteristic for the rural counties of the United States, such as the number of children under five to 1,000 women of childbearing age for some censal year. If we define as a "rural county" every county which has 100 percent of its population rural and which is nonadjacent to a "metropolitan" or an "industrial" county, the limited universe which we are studying consists of around 1,000 such counties. If by the use of Tippett's random sampling numbers or some other such scheme we draw a sample of 20 rural counties, or 2 percent of the universe, the formulas developed for an infinite universe may be applied to estimate the unreliability of our estimates of parameters of the limited universe. Now if we increase the size of our sample to 200 rural counties, we are including one fifth of the universe, which is an appreciable portion of it, we are about at the borderline where we should begin to use formula (8), for instance, in computing the standard error of the mean. Evaluating in part for formula (8), we have

$$\hat{\sigma}_{\bar{x}_{'}} = \hat{\sigma}_{\bar{x}} \sqrt{\frac{1,000 - 200}{1,000}} = \hat{\sigma}_{\bar{x}} \sqrt{\frac{800}{1,000}} = \hat{\sigma}_{\bar{x}}\sqrt{.8} = .894\hat{\sigma}_{\bar{x}} \qquad (9)$$

Now if we increase the size of the sample even more, the coefficient .894 in the rightmost member of expression (9) becomes smaller and smaller, and when the sample is of 1,000 units, the entire universe, the expression becomes zero, meaning of course that there is no unreliability to our estimate. In fact, the value of $\bar{X}$ is no longer an estimate, but it is the actual measured universe parameter, μ.

If we restrict our procedures to the methods of descriptive statistics, the case is clear and simple; we have secured one or more descriptive measures for the finite universe we are interested in, and sampling and tests of significance have no meaning or application to the problem. This is the point of view of certain statisticians on the problem.[6] There is another point of view, however, not too well clarified as yet, which we shall attempt to present.

Tests of significance when all units in a limited universe have been observed. For the sake of making the problem more meaningful, let us anticipate some of the concepts and procedures of the next part of this text—those which are a part of the statistics of relationship. Let us

---

[6] *Ibid.*, pp. 521–522.

suppose that instead of investigating only fertility ratios in the rural counties, we investigate simultaneously fertility ratios and the index of level of living in the counties. In studying these two distributions simultaneously, we use the coefficient of correlation, r, to describe the intensity or degree of association between the two characteristics. If in the sample of 20, we compute the r, we can estimate from this the value of ρ, the coefficient of correlation in the universe, and compute the standard error of $\hat{\rho}$ and its confidence limits. Again as the number of cases in the sample approaches the number of units in the universe, the standard error of $\hat{\rho}$ approaches zero, and when $N = P$, $\hat{\rho} = \rho$, with no error involved in the estimate. Now this measure of degree of association or relationship between two characteristics is an important type of finding in sociological research, since it may under certain conditions be used as a first step in unravelling a causal nexus of factors. The immediate point at issue becomes this—is a parameter which has been computed on the basis of observation of all the units in a finite universe to be used solely as a historically descriptive measure of some aspect of the distributions of characteristics in that universe? If so, there is no meaning to the standard error of the parameter, and we can make no statistical tests of hypotheses relating to the parameter.

Yet, one continually finds in sociological research reporting, as in other fields, standard errors computed for coefficients of correlation or other measures on a series of units comprising all of a kind existing—for instance, on all 48 states, on all metropolitan areas, or on all census tracts in a certain city as of a certain date. Furthermore, on the basis of the standard errors, tests of significance are made and interpretations of them often include reference to a "true" value or a "real" difference. In spite of this practice, it is seldom in research reporting, or even in articles or texts on statistical methods and interpretation, that one finds much elucidation on what is meant by such tests or their interpretation.

The universe of possibilities. Let us examine what possible meanings standard errors and statistical tests of hypotheses could have in such cases where all units in a finite universe have been measured or enumerated. First, there must be some superuniverse to which the hypotheses, and such terms as "true" or "real" value refer. This superuniverse must be a universe from which our finite universe can be considered a random sample. It has been defined as an unlimited or infinite hypothetical universe of possibilities—the universe of all the possible finite universes that could have been produced at the instant of observation under the conditions obtaining.[7] It is therefore only an imagined possibility, and

---

[7] The clearest exposition of the concept of the universe of possibilities is found in "Sociology and Sampling," by Samuel A. Stouffer, in L. L. Bernard (ed.), *Fields and Methods of Sociology* (New York: Long and Smith, 1934), pp. 476–487.

whether or not one wishes to utilize the concept is still at the discretion of the individual research worker.

Meaning of the universe of possibilities in experimental work. Since the construct of a hypothetical universe of possibilities has proved useful and fruitful in other fields of research, let us look at its application in one of them, that of agricultural experimentation. The ideal of any experimental setup is to control all conditions save the factors being studied and then in terms of correlation coefficients and other measures developed in the statistics of relationship to describe the relationship existing between the one or more "independent" or "causal" factors and the "dependent" or "effect" characteristic. With modern methods of design of experiment and analysis of data, the effect of several "independent" factors such as fertilizer, variety of plant, spacing, and other controllable factors on the "dependent" characteristic, or yield, may be studied simultaneously. We wish to consider the simplest case, however, where the relationship between only one independent fact such as amount of fertilizer and the dependent factor, yield, is being investigated. Let us suppose that for a certain range in amount of fertilizer used, there is found to be a linear correlation between amount of fertilizer and amount of yield described by the coefficient r based upon an experiment involving N observations, one on each of N plots. As far as the methods of descriptive statistics go, the value of r precisely describes the degree of association between amount of fertilizer and amount of yield for this particular group of N plots for one season. But no experimentalist trained in statistical methods would stop at that point; he would test his observed r to see if it were significantly different from zero. More exactly, he would test the hypothesis that his sample of N observations with an observed correlation of r could have been a randomly drawn sample from a universe where fertilizer and yield were uncorrelated, or where $\rho = 0$. If the test showed that the probability of observing a sample of N cases with a correlation as great as r from such a universe was very small, the experimenter would reject the hypothesis $\rho = 0$ and implicitly or explicitly affirm the hypothesis that $\rho > 0$. He would interpret his results to mean that he had observed a *significant* coefficient of correlation in his sample of N observations—significant in the sense that the observed r *signified* a ρ different from zero in the universe from which the sample had been drawn.

So far the procedures are straightforward, but the point we are interested in is more elusive. It is in answering this question. What is the precise definition of the universe concerning which the hypothesis was tested and from which the set of N observations may be considered a random sample? The experimentalist defines the universe as all the possible results that would be obtained by repeating the experiment under

identical conditions an infinite number of times. One may ask how he knows he is justified in assuming that his one set of experimental data may be expected to yield values of parameter estimates which obey the laws of a sampling distribution theoretically derived for the mathematical model of random sampling. The answer is that a great mass of empirical proof has shown this to be the case. In fact, experimental techniques are often submitted to testing in what is known as uniformity trials where the criterion is conformity of their results to such sampling distributions.[8]

Two alternative interpretations of chance factors. Granted that experimental technique has not been perfected in any science to the extent where absolute identity of conditions can be obtained in repeated experiments, empirical evidence shows that in some fields it can be secured to the degree that only such variations as would be expected in random sampling are uncontrollable. The next question to be answered is—To what are these residual variations due? Their distributions resemble the distributions we would expect if there were actually random sampling from an infinite universe, and therefore they are often said to be due to "chance" or to "random sampling variation." Again we reach the matter of defining "chance" and any attempts lead immediately into theories of causation. It may be that there is rigid and specific determinism of each event which occurs, and that our ascribing the residual variations we cannot explain to "chance" is merely a way of placing the limits of present day scientific knowledge. In such an interpretation the level at which knowledge stops may be different for different fields of science—it certainly reaches to subatomic levels in modern physics, whereas in certain aspects of agricultural biology it may stop with factors external to the living organism. If such an interpretation be accepted, we can expect to see advances in science reduce the unexplained variation we now ascribe to "chance" toward the limiting value of zero when knowledge is perfect, whether or not the limit ever be achieved. Just what interpretation is to be attached to the fact that the variations we cannot at any one stage explain are distributed according to the expectation of random sampling is not completely clear. The usual explanation for the close correspondence between observation and theory is that the factors labeled "chance" are numerous, each relatively small in importance, and independent of each other, for it can be theoretically deduced that this sort of situation would lead to such variation as is actually observed.

Certain developments in modern physics have implications for a less rigid determinism of the occurrence of events. Such interpretations of causation admit the impossibility of predicting unique events, because of the fact that only group averages are determined, with dispersion of

---

[8] George W. Snedecor, *Statistical Methods Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), p. 214.

individual events around the average in distributions similar to those described by random sampling theory. Indeterminacy according to such an interpretation of causation is not due to a mere limitation of the state of knowledge, but is an inherent property of the behavior of events, which are determined only in terms of group averages. By this interpretation, probability is a central concept of all treatments of causation, and the distributions expected from random sampling are a basic and fundamental aspect of the description of ultimate facts of causation. The implication of this interpretation is that no matter how advanced the state of knowledge becomes, scientific prediction will always have to be done in terms of probabilities less than unity.

The experimentalist's practical interpretation. The meaning and implications of these two interpretations of causation should be faced, but one does not have to "believe" in one or the other to make adequate interpretations of statistical analysis. To get back to the case of the agricultural biologists, his interpretation of the "significance" of his observed correlation coefficient is usually oriented to a practical situation. On the practical level, the fact that his r is "significantly" different from zero and (let us assume) positive means that he has grounds for confidence that a greater amount of fertilizer on this particular variety of plant will cause a higher yield from it, not only in a rigidly controlled experimental situation, but under actual farming conditions also. His confidence is based upon the approximate correspondence between the variation observed under experimental conditions controlled to the extent that identical conditions are approximated, and the variation of a distribution theoretically expected from random sampling of an infinite universe. Since the results of the experimental and mathematical models correspond, he can express his expectation of variation in his experimental situation in terms of the probabilities deduced from the mathematical situation, realizing that the approximate nature of the correspondence limits to some extent the literalness with which he must interpret precise probabilities. The validity of his practical predictions is not determined by what he interprets the "chance" forces to be, so long as he can describe the results of their operation. Thus, the experimentalist utilizes the statistical analysis of his results to explain phenomena (insofar as measuring degree of association where causation is imputed may be called explanation), to predict future phenomena (for instance, more yield from more fertilizer), and thereby to provide a scientific basis for the prediction and control of phenomena. This is an exaggeratedly simple case, of course, for purposes of illustration to contrast the situation of the experimentalist with that of the mere observer who has little or no control over the phenomena he is studying.

Interpretation in the nonexperimental situation. Sociologists, especially those in the field of social psychology, are developing experimental

techniques for attack on certain of their problems where they have the advantage of more or less control over relevant factors. In general, however, this is not the case, and we are concerned here with the situation where the sociologist can only observe at one time measures or attributes of a series of units. We are concerned especially with the case where the observations are made for all of a series of units as of a certain time, as, for instance, on all of the rural counties in the United States for 1930. Let us suppose, to continue the illustration begun earlier and interrupted, that for 1930 one observes a negative correlation coefficient of value r between fertility ratio and level of living for all rural counties in the United States. With the same formulas and procedures used by the experimentalists, the sociologist tests the significance of his observed r and finds it "significantly" different from zero. What does this test mean to the sociologist? In any test of significance there is a testing of some hypothesis about a universe from which the set of observations (a limited universe itself in this case) may be considered a random sample. That is, the logical structure of a superuniverse and the variation expected in random samples from it is the same for the observer sociologist as for the experimentalist.[9] Imagining any experiential counterpart of the logical model is a more difficult matter, however. It is easy enough for the experimentalist to imagine repeated experiments under identical conditions, whether or not he can actually perfect his technique to the degree that he can reproduce conditions identically. His universe of possibilities can, therefore, be put into meaningful terms; it can at least be imagined, even if it cannot actually be produced. It is not so easy for the sociologist to imagine a set of observations repeated under conditions identical with those of one date. The fact of change in social and cultural phenomena renders unrealistic any conception of identical repetition of the complex of factors conditioning characteristics such as fertility and level of living. The concept of the universe of possibilities—that is, all possible sets of measures on fertility and level of living that could possibly be produced in the thousand rural counties of the United States under conditions exactly similar to those of 1930—the concept has neither a realistic

---

[9] This sentence deserves certain qualification. The observational situation often differs from the experimental in that other relevant factors are not constant for all of the observed units. The concept of the dynamic universe as explained by Thomas C. McCormick involves the requirement that each observed unit in the sample have the same probability of possessing an attribute or a given value of a variable. Such homogeneity can sometimes be approximately obtained in the observational situation by successive subclassification with respect to all known influencing factors as is illustrated by McCormick in "Sampling Theory in Sociological Research," *Social Forces*, 16 (October 1937), pp. 67–74. Or a statistical substitute for homogeneity can sometimes be obtained by partial correlation techniques. But in neither the experimental nor the observational situation can absolute conformity to the criterion of homogeneity be obtained, and the essential difference between the two situations seems to be one of degree of approximation to the criterion. How gross this approximation may become before the validity of generalizing to the dynamic universe of possibilities is completely vitiated is an unsolved question.

counterpart nor a readily imaginable counterpart. To what, then, does the variation expected from random sampling from such a universe of possibilities correspond? Only a feat of imagination involving an infinite prolongation of a present moment, where conditioning factors remain the same but "chance" factors continue to produce random variation can supply the answer. With this done, the observer sociologist along with the experimentalist still faces the problem of interpretation of the chance variation—with the alternatives of ascribing it to the present limitations in knowledge or to the statistical nature of the occurrence of events.

Other more realistic models may occur to the reader, but so far none has been proposed which is satisfactory. It has been suggested that the limited universe of measures on all of a series of demographic units as of a certain date be considered a sample in time; that the random variations of sampling from a superuniverse have their counterpart in the fluctuations which would be observed if we made observations on successive days, or for successive years, while the general influencing conditions would not have altered appreciably. It is evident, however, that such successive fluctuations would not be independent, nor could they be thought of as being produced by forces independent of each other, and therefore they would not be expected to have the same distribution as those produced by chance factors in random sampling (as in fact can be shown to be the case). This approaches the problem of the economist in the interpretation of fluctuation in a time series.[10]

Another suggestion is that the measures on demographic units may be conceived of as one of an infinite set of such measures secured by dividing the total area surveyed into different series of areal units by shifting of boundaries under certain conditions of contiguity and uniformity of size. The matter of the arbitrary nature of the "lumps" in which our demographic information is secured, and the possible variations to be expected by recombining the information into different lumps has not been explored adequately. While the matter needs attention, it is probably not the

---

[10] Excerpts from a noted economist's work on certain time series will illustrate this:

Now time series, especially those relating to social and economic phenomena, are likely to violate in a marked degree the fundamental assumption which underlies the use of the methods sketched above, namely, that not only the successive items in the series but also the successive parts into which the series may be divided must be random selections from the *same* universe. Time series are, in fact, a group of successive items with a characteristic conformation. Such series . . . cannot be considered as a random sample of any definable universe except in a very unreal sense. Nor are the successive items in the series independent of one another. . . . The fact is that the "universe" of our time series does not "stay put," and the "relevant conditions" under which the sampling must be carried out cannot be re-created. . . . It is clear, then, that standard errors derived from time series relating to social and economic phenomena do not have the same heuristic properties that they have, or are supposed to have, in the natural sciences.

Henry Schultz, *The Theory and Measurement of Demand* (Chicago: University of Chicago Press, 1938), pp. 214–215.

answer to the search for a realistic counterpart of the universe of possibilities and to the random variation expected in samples from such a universe.

Reasons for using the construct of a hypothetical universe. At present the sociologist must face the fact that the postulated, hypothetical, infinite universe of possibilities, concerning which he tests hypotheses to establish the "significance" of his results, is merely a logical structure, for which he can offer no real counterpart in his research situation. Then what is the utility of such a construct and of the tests of significance based upon it? The answer to this question is not perfectly clear at the present stage of the application of statistical methods to sociological research. A case for the use of such a construct may be made on the basis of the following considerations, however. The amount of variation due to chance factors expected in statistics of random samples can be used as a standard, against which we evaluate variation observed in two different samples or hypothesized between a universe of possibilities and an observed sample. This affords a criterion for differentiating between variations which may be regarded as accidental or fortuitous, and those which cannot be so regarded. Another consideration in favor of the use of such a construct is that it has proved fruitful in other fields of research and therefore deserves a fair trial in sociological research, although interpretations of other fields cannot be slavishly imitated since the situations are so different. Finally, there is another reason for trying to generalize to the superuniverse of possibilities, which can only be suggested, since the concept has not been well clarified by those engaged in sociological research. It is based on the premise that there is a stability, a regularity, an orderliness in the occurrence of sociological phenomena, even though it is dynamic and ever-changing, and that one task of developing a scientific sociology embraces the description and formulation of the stable and regular, though dynamic, relationships underlying two or more series of phenomena. We have stated previously that the fact of differences in geographic location, culture, and time seems to preclude the possibility of developing any truly universal laws, or descriptions of relationships among series of social phenomena which would be valid for all times, places and cultures. Therefore, our goal in developing a scientific sociology is necessarily limited in the description of relationships. Yet, somehow, there seems to be a place for the sifting from sets of observed measures of relationships the irrelevant variations which particularize them as unique, in a search for meaningful relations, impermanent and varied with location though they be. This goal is so far short of those of the physical sciences that it may be misleading to use the term scientific in our field. And yet at the present stage, the goal seems to be the only realistic one. It seems also that the transition from analysis of observed data on a finite universe by the methods of descriptive statistics to the use of the methods of inductive

statistics in inferring information about the universe of possibilities—be it only a logical construct—is one approach to this limited goal.

SUGGESTED READINGS

Blankenship, A., *How to Conduct Consumer and Opinion Research*, 2d ed. (New York: Harper, 1945).

Deming, William Edwards, *Some Theory of Sampling* (New York: Wiley, 1950).

Deming, W. Edwards, "Some Criteria for Judging the Quality of Surveys," *The Journal of Marketing*, XII (October 1947), pp. 145–157.

Hansen, M. H., and Deming, W. E., "On Some Census Aids to Sampling," *Journal of the American Statistical Association*, 41 (1943), pp. 353–357.

Mosteller, Frederick, and others, *The Pre-Election Polls of 1948: Report to the Committee on Analysis of Pre-election Polls and Forecasts* (New York: Social Science Research Council, Bulletin 60, 1949).

Neyman, J., "Contribution to the Theory of Sampling Human Populations," *Journal of the American Statistical Association*, 33 (1938), pp. 101–116.

——. "On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97 (1934), pp. 558–625.

Stephan, Frederick F., "Representative Sampling in Large-Scale Surveys," *Journal of the American Statistical Association*, 34 (1939), pp. 343–352.

Yates, Frank, *Sampling Methods for Censuses and Surveys* (New York: Hafner, 1949).

Yule, G. Udny, and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), Chaps. 16–19, 23.

CHAPTER 18

∿∿∿
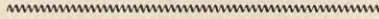
Sampling in Social Research: Application in Surveys

Purpose of chapter. A general text on social statistics cannot cover fully the details of the advances in application of sampling methods to surveys that have taken place in recent years. The purpose of this chapter is quite modest. A brief history is given of the kinds of sampling methods that have been used in social and economic surveys in the United States. The general nature of recent advances in the use of sampling in surveys is set forth. An account of the sampling resources available to social research workers—both in the nature of materials and of consulting services—is provided. The growing importance of the use of sampling by governmental agencies in the collection and processing of statistics that are used by sociologists and other research workers is pointed out with general descriptions of illustrative types of sampling methods used. And, finally, a concrete illustration is provided of the use of sampling resources of federal agencies in a survey of two economic areas within a state.

BRIEF HISTORY OF SAMPLING METHODS USED IN SOCIAL AND ECONOMIC SURVEYS IN THE UNITED STATES

One of the characteristic features that has distinguished American sociology from European has been its greater emphasis on empirical research—on observing and analyzing social phenomena first hand through field studies or other arrangements for obtaining information on human beings, their behavior, and their institutions. This has meant that surveys or other types of first-hand collection of data have occupied an important role in the development of methods of social research in the United States.

Some early approaches. Prior to the 1940 decade social surveys in the United States used sampling methods that would not be considered satisfactory according to present-day standards, and yet there were some

field studies going on that might be considered precursors to modern sample surveys. In the 1920's rural sociologists and agricultural economists in the state colleges of agriculture began to receive research funds from federal sources which permitted the undertaking of field studies. Hit-or-miss methods were generally employed in the selection of rural families, communities, or farms that were to be included in the survey, but the objectives of sampling were becoming formulated and gaining some acceptance in rural social research. The need was becoming recognized for the development of techniques for the selection of a fraction of the phenomena of interest in such a way that the fraction "represented" the larger universe of such phenomena. For example, in the early 1920's Charles Galpin in the Bureau of Agricultural Economics initiated annual mail questionnaire surveys on the movement of the farm population, using the lists of farmers known as crop reporters. These lists had been prepared by agricultural statisticians in each state to obtain data by mail on which to base crop and livestock estimates. They were compiled by various methods, and the extent of bias in considering the lists as a sample of all farmers in the United States was unknown. Likewise the additional bias due to the fact that only a fraction of the farmers returned the questionnaires was unknown. Nevertheless, this was the first use in the United States of a national sample survey of a segment of the population (those living on farms) to provide the basis for current estimates of population and migration.[1] Also, Dr. Galpin stimulated field studies involving local surveys by rural sociologists in the land grant colleges, often making funds or personnel available to assist in such studies. An example of one of these early studies in Carle C. Zimmerman and Carl C. Taylor's *Rural Organization, a Study of Primary Groups in Wake County, N. C.* (North Carolina Agricultural Experiment Station, 1922).

In the urban field the subjects of household or population sample surveys of cities before 1930 included income, expenditures, unemployment, and health.[2] Frederick F. Stephan considers noteworthy as an advance beyond previous practice the survey conducted in Columbus, Ohio, in 1922 by Frederick S. Croxton and Mary Louise Mark. Three districts within the city were selected by representatives of business and labor as a "fair sample of the wage-earning population of the city" for canvass in the survey. Toward the end of the decade committees of the Social Science Research Council began to give attention to sampling. The work of A. L. Bowley in England was given more consideration by Ameri-

[1] For further description of the methods of developing the annual farm population estimates from these surveys, see "Farm Population Estimates" Census-BAE, No. 16 (Washington; Bureau of the Census and Bureau of Agricultural Economics, 1952).

[2] For more detail on these surveys see Frederick F. Stephan, "History of the Use of Modern Sampling Procedures," *Journal of the American Statistical Association*, 43 (1948), pp. 12–40.

cans after Margaret H. Hogg, who had worked with Bowley, came to this country near the end of the decade and urged that random procedures be used rather than judgment in selection of samples.

Developments in the 1930's. It is somewhat ironical that the greatest depression in the history of the United States directly and indirectly served as a powerful stimulus in the development and application of sample survey methods. One important factor was the need for information on the number and types of the unemployed.[3] As the federal government entered into the field of relief and aid to various special groups, there was need for many types of information to be used to guide the policy of these programs that could be obtained by sample surveys. At the same time the federal work relief program needed to provide useful employment to workers who could be utilized in research projects, and, consequently, it set up many research projects, often in cooperation with state or local agencies, that involved sample surveys.

The number of sample surveys undertaken in this program was too great for the methods to be presented in detail here. The direction of the changes in sample survey methods that were applied was, in general, away from purposive or subjective selection toward objective methods involving systematic or random selection. Also, toward the end of the decade exploration began of various types of stratification and of the various types of sampling units (county, township, block, clusters of families, individual households, etc.) to be used in surveys.

Although the 1930 decade witnessed substantial progress in the application of sampling in social surveys, we wish to stress that perfection was not attained and that one should not at present follow the methods used in the 1930's as models for survey design. In general, it was the beginning of a period of transition, but the transition was far from completed in the decade. Because numerous graduate students in the social sciences at that time were connected with various Works Projects Administration surveys (many got their field work experience and the data for their theses from such surveys), many social science professors and research workers of the present hold to certain ideas about sampling that were current in the late 1930's but that are now outmoded. Several of these will be mentioned.

There was some adherence to the view that *representativeness* must be assured by methods other than proper design of a sample and leaving it to chance to determine which units actually fall in the sample. This led to the type of sample presented in *Rural Regions of the United States*. After a

---

[3] See John D. Durand, "Development of the Labor Force Concept," Appendix A in Louis J. Ducoff and Margaret Jarman Hagood, *Labor Force Definition and Measurement: Recent Experience in the United States* (New York: Social Science Research Council, Bulletin 57, 1947).

careful job of stratifying the counties of the United States into a number of relatively homogeneous subregions, a national sample was selected by choosing one county (in special cases, more than one) from each of the subregions. The choice of the county was done objectively after a criterion of selection had been subjectively chosen. The criterion involved choosing the county that was closest to the median for the subregion on three control items—plane of living, fertility ratio, and percent of farms producing less than $1,000 gross income.[4] The flaw introduced by this procedure is that those counties which departed from the average for the subregion containing them had no chance to fall into the sample. The nature and amount of bias so introduced is unknown. For certain purposes such a sample might provide a good estimate of some characteristic for the universe, but sampling theory cannot provide the basis for determining the probability that the estimate will depart from the universe value by some specified amount.

An extension of the adherence to this view is that in any type of stratified sample the units chosen in one stratum must be *representative* of that stratum or that for any geographic division within the area which is the universe the sample units *should be representative* of the division.[5] In the first case the design of a stratified sample may call for one unit or a number of units to be drawn from each stratum. The precision with which the sample units can provide an estimate for the stratum depends on the number drawn from the stratum, and we have no right to expect the precision to be high from a sample of one or a few units. Similar reasoning applies in the second case. A sample of 200 counties in the United States might have four counties in a given state which provide a very poor sample for that state, even though at national and even regional levels the sample might be entirely satisfactory. Criticisms of national samples have often been made unjustly on the basis that the county or counties included in a certain state were very atypical, when this fact provides no basis whatsoever for a negative appraisal of the sample.

Another view in the 1930's was the great emphasis on *validation* of the sample. Because many of the methods of sample selection used were defective, it was considered of prime importance that the representativeness of the sample be checked or validated—before it was used, if possible. It is still considered desirable that any results from sample surveys be checked with independent data if such exist. But the emphasis has totally altered with the greater emphasis now on proper design and selection of a

---

[4] This is a slight oversimplification of the methods of selecting the sample. For a fuller account see A. R. Mangus, *Rural Regions of the United States*, (Washington: Government Printing Office, 1940), pp. 92–95.

[5] For fuller discussion of this subject, see Earl E. Houseman, "Design of Samples for Surveys," *Agricultural Economics Research*, Vol. 1, No. 1, (Washington: Bureau of Agricultural Economics, January 1949).

sample that will yield estimates with a known range and probability of error. The principal assurance of the *validity* of the sample now is the method of design and selection of the sample combined with high quality field work, construction of the questionnaire, and the drawing of conclusions in light of knowledge of sampling error and possible biases.

Developments in design of samples. There have been a number of developments in the methods of designing samples for surveys since the middle of the 1930's. These have included the introduction of several steps in the sampling process, such as drawing first a sample of primary sampling units and within these drawing secondary sampling units, each of which may include only one or a cluster of households or other types of units for which a schedule is to be taken. The reasons for this general type of sample design stem largely from considerations of administration and cost of field operations. When there is sampling in two or more stages, precautions must be taken to maintain the requirement that each unit has the desired probability of coming into the sample. In two-stage sampling the probability that a unit will come into the sample is the product of two probabilities—the probability that the primary sample unit will be drawn times the probability that the unit will be drawn from the units in the primary sampling unit. In surveys of households or farms in the United States primary sampling units are often counties or groups of two or more adjacent counties. For example, in designing a national sample of farm-operator households, one might group the counties of the United States into strata, select one county as a primary sampling unit from each stratum, then select certain proportions of segments (small areas) in the sample counties, and then include in the survey all or a subsample of the farm-operator households in the sample segments. In carrying out the selection of counties, it must be certain that each farm in the United States is given a known probability of getting into the sample. A common way of handling this is to form the strata in such a way that each stratum includes approximately the same number of farms. Next, we use a method that will give each county a probability of entering the sample proportional to its number of farms. A subsampling fraction is then selected to use within the county so that each farm operator will have an equal chance of being part of the sample. A convenient procedure for handling this is to list the counties in some order (any order) with the number of farms opposite each county name. Now, total the number of farms (preferably by an adding machine), taking a subtotal after each county's number of farms is added. If the counties are average size, the total number of farms in a stratum of 100 counties will be a little under 200,000. Next, devise a random procedure for entering a list of six-digit random numbers and take the first such number that is less than or equal to the number of farms in

the stratum. Then, identify the county enclosed by subtotals which include the random number, and select this county for the sample. The subsampling fraction from the selected county then should be proportionate to the reciprocal of the number of farms in the county. Thus, the subsampling fraction will be smaller if a large county is selected or larger if a small one is selected.

We have described one step in conducting a sample design to illustrate the type of improvements in design that have occurred since the 1930's. We shall mention another type involving the estimating procedure in the section on the Current Population Survey of the Census Bureau. But we wish to make clear that the field of design of samples for surveys has become an area of specialized knowledge that cannot be systematically covered in a brief treatment. Research workers without specialized training in sampling are urged to consult sampling specialists if they have an occasion to undertake a sample survey.

The public opinion polls and many market research surveys have relied mainly on design of samples involving quota sampling referred to briefly in the preceding chapter. Because we do not recommend quota sampling generally to social research workers planning surveys, we are not treating this subject or the controversy over quota versus probability sampling.[6]

Development of area sampling.   One of the first major landmarks in the development of area sampling was a survey in Iowa conducted in the winter of 1938–1939.[7] The results were so promising that the method was soon extended to other areas of the United States.[8] Another landmark in area sampling was the development by the Iowa Statistical Laboratory, the Bureau of Agricultural Economics, and the Bureau of the Census, of the "Master Sample," which was used in the 1945 Census of Agriculture.[9] The original work was focused on methods for sampling farms and farm population, but the scope was expanded to cover the sampling of households or population in both rural and urban areas. The Bureau of the

---

[6] The following references are suggested for treatment of this subject: Frederick Mosteller and others, *The Pre-election Polls of 1948: Report to the Committee on Analysis of Pre-election Polls and Forecasts* (New York: Social Science Research Council, Bulletin 60, 1949); Hochstim and Smith, "Area Sampling or Quota Control?—Three Sampling Experiments," *Public Opinion Quarterly*, 12 (Spring 1948), pp. 147–155; W. Edwards Deming, "Some Criteria for Judging the Quality of Surveys," *The Journal of Marketing*, XII (October 1947), pp. 145–157.

[7] Raymond J. Jessen, *Statistical Investigation of a Sample Survey for Obtaining Farm Facts* (Ames, Iowa: Agricultural Experiment Station, Research Bulletin 304, 1942).

[8] Raymond J. Jessen and Earl E. Houseman, *Statistical Investigations of Farm Sample Surveys Taken in Iowa, Florida, and California* (Ames, Iowa: Agricultural Experiment Station, Research Bulletin 329, 1944).

[9] Arnold J. King and Raymond J. Jessen, "The Master Sample of Agriculture. I, Development and Use; II, Design," *Journal of the American Statistical Association*, 40 (March 1945), pp. 38–56.

Census has made extensive use of area sampling in its current population surveys and in its monthly retail trade survey.[10]

A central problem posed in the development of area sampling was what kind and size of area to use in designing a sample of farms in a state? A county, a township, a section, or a fraction of a section? From sampling theory it is known that the reliability of estimates is greater (for a given number of farms in a survey) when the areas are small, but it is obvious that this is more expensive for conducting field surveys than when large clusters of farms are surveyed. For example, for a state sample of 800 farms it is much cheaper to conduct the survey in 40 areas, each containing about 20 farms, than in 400 areas each containing about two farms, although estimates based on data from the latter would have a smaller sampling error.

The work at Ames and elsewhere led to the adoption of segments (small areas) for general farm or farm population surveys that average from four to eight farms, varying in size somewhat for the different regions of the country. Such segments have been delineated on maps for each county of the United States and numbered in a serpentine fashion within a county. These constitute the "Master Sample" materials for the open country and places of less than about 100 population. Somewhat similar methods applied to maps for cities, smaller incorporated and unincorporated places of 100 or more population provide a complement that permits the materials to be used for a sample of the entire population. Given specifications as to the type of residence of the population to be covered—such as total open country, or all rural—and as to the rate of sampling desired—such as 1 percent—these materials can be used to draw a sample and provide the maps for field work for a survey to cover farms or population (or farms of a particular kind if frequency of occurrence is not too low) for any county or group of counties in the United States. The largest scale single use of area sampling made to date was in connection with the 1945 Census of Agriculture, in which an area sample of about 300,000 farm operators throughout the United States were asked additional detailed questions.[11] In the aggregate the repeated monthly Current Population Surveys of the Census represent an even larger use. In addition, many federal agencies, state colleges, and commercial institutions have used area sampling in surveys by either developing their own approach or taking advantage of the "Master Sample" sampling materials.

---

[10] See *History, Operations, and Organization of the Bureau of the Census*, Vol. I (Washington: Bureau of the Census, 1946), pp. 29–45.

[11] Arnold J. King and Raymond J. Jessen, "The Master Sample of Agriculture. I, Development and Use; II, Design," *Journal of the American Statistical Association*, 40 (March 1945), pp. 38–56. Also see "Census of Agriculture, 1945: Special Report," 1945 Sample Census of Agriculture (Washington: Bureau of the Census, 1947).

Development of techniques of field operations.   The nature of developments in techniques of field operations will be touched only briefly, although they are of great importance in the conduct of social surveys. In large-scale surveys new visual and auditory aids have been utilized in the training of survey enumerators. There is more rigorous pretesting and quality testing to appraise the results of surveys. Detailed records now provide better bases of cost estimates for surveys being contemplated. The conduct of field operations in large-scale surveys has, like sampling, become a specialized field, but social research workers undertaking smaller surveys can benefit from the experience in field operations of larger organizations. The following references are suggested:

Monographs that it is hoped will be issued on the experience in field operations of the Seventeenth Decennial Census.

Pauline V. Young, *Scientific Social Surveys and Research* (New York: Prentice-Hall, 1946).

Albert Blankenship, *Consumer and Opinion Research* (New York: Harper, 1943).

Albert Blankenship (ed.), *How to Conduct Consumer and Opinion Research* (New York: Harper, 1946).

C. W. Churchman, Russell L. Ackoff, and Murray Wax (eds.), *Measurement of Consumer Interest* (Philadelphia: University of Pennsylvania Press, 1947).

SAMPLING RESOURCES NOW AVAILABLE TO RESEARCH WORKERS IN THE UNITED STATES

Theory and literature.   In contrast with the WPA days, sampling theory and literature on its application to social surveys is considerably more available now. However, there is too great a time lag between the new developments in application of theory to design of sample surveys and the publication of literature that would enable the individual research worker (or small institution) to apply the new improved methods to his own sample survey problems. This arises partly from the fact that those most sophisticated in sampling become aware of more and more problems and complexities that have to be met in actual sampling situations and they recognize that they cannot offer simple, foolproof patterns to be followed blindly by a person who is not trained in at least the elementary principles of sampling. Two general works on the subject have appeared since the first edition of this text, which are highly recommended to the reader:

Frank Yates, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin; New York: Hafner, 1949).

William Edwards Deming, *Some Theory of Sampling* (New York: Wiley, 1950).

Sampling consulting service. At a number of the leading colleges and universities in the United States sampling consulting service is available from specialized personnel in departments of statistics or in research institutes. Even in these more favored institutions, however, full utilization of consulting service is often not made by individuals or groups planning projects that involve sampling. And in many educational institutions which do not have such consulting resources, the individual research worker may not know where to turn. A great deal of sampling consulting service is supplied on request by the institutions, both private and governmental, which have specialists in sampling methods, but a great need exists for an increase in the availability and the utilization of sampling consulting service. Also, social research workers should understand that the time sampling consultants can be of the greatest aid is during the stage of design of their projects. Too often the consulting service is sought only after the sample survey has been made, and the consultant can only offer suggestions for how to patch up a very bad job.

Sampling maps and materials. The three places where sampling materials are on file for every county of the United States are the Statistical Laboratory of Iowa State College at Ames, Iowa, and the Bureau of Agricultural Economics and the Bureau of the Census in Washington, D. C. At each of these places there are materials that permit ready sampling of the open-country territory of nonmetropolitan counties. A county highway map showing culture has been divided into "count units," set up with the objective of containing 10 to 20 farms in most parts of the country. On the map the count units have been numbered, and the estimated number of farms in each is recorded. The farms have been added with a subtotal taken for each count unit, as described above, to facilitate the process of drawing a sample. There are also listings of unincorporated and incorporated places with a population estimate for each. At Ames and the Bureau of Agricultural Economics there is no comprehensive collection of maps for the unincorporated centers, the incorporated villages and the urban places. When either of these agencies draws a sample for other than open-country territory, it is usually necessary to secure additional maps or aerial photographs. The Bureau of the Census has the most complete collection in existence of maps of cities in the United States, including the Sanborn maps for larger cities,—most cities of 25,000 or more population—and a few smaller cities. The Census Bureau also has certain published statistics for blocks already on the maps for larger cities to facilitate the drawing of urban samples.

Each of these agencies uses its sample materials for drawing samples and preparing maps for its own surveys or for surveys it makes in cooperation with state or other agencies. The Bureau of Agricultural

Economics has used its materials to develop samples for other agencies within the Department of Agriculture and for land grant colleges. The Ames Laboratory and the Census Bureau have arrangements whereby they can provide sampling services to any agency or individual at cost. Because the Census Bureau has negatives of the maps used for the open-country samples, reproduction of these by the ozalid process is quite inexpensive.

If a private research worker or institution wishes to use the sampling resources at Ames or the Census Bureau, the steps involved are as follows. Through personal visits or correspondence with one of these agencies the type of sample needed is discussed. If the research worker wishes one of these agencies to draw the sample and prepare the necessary maps, he specifies his requirements and is given a cost estimate. The costs of open-country sampling materials are usually moderate because the material is in form for low-cost reproduction. In a majority of cases where surveys of farms or the farm population are made the costs are probably no more than 5 percent of the total survey costs and often less. Block maps and statistics for large cities are available at still lower costs. Costs for small- and medium-sized cities and suburban areas are higher. If the research worker decides to utilize the facilities, he places an order and pays the cost involved. The facilities and arrangements mean that persons in the United States engaged in social research work that involves sample surveys have at their command for relatively small cost sampling materials and advice from nonprofit public agencies that are known throughout the world for the expertness of their sampling personnel.

EXAMPLES OF USE OF SAMPLING IN DEVELOPMENT OF SOCIAL AND ECONOMIC STATISTICS BY FEDERAL AGENCIES

There is wide use of sampling by federal agencies in the production of many types of statistics. We will not attempt to cover this whole field but merely to give three selected examples: (1) a periodic sample survey; (2) a use of sampling in census enumeration; and (3) a set of special-purpose sample surveys.

The Current Population Survey of the Census Bureau. In 1942 a monthly survey known as the Monthly Report on the Labor Force was transferred to the Bureau of the Census and has been continued there with various modifications ever since. The original sample was designed by Lester R. Frankel and J. Stevens Stock,[12] but this will not be described since the sample was redesigned in late 1943 by Morris H. Hansen and

[12] Lester R. Frankel and J. Stevens Stock, "On the Sample Survey of Unemployment," *Journal of the American Statistical Association*, 37 (February 1942), pp. 77–80.

William N. Hurwitz of the Bureau of the Census.[13] Since then there have been further modifications to the sample design.

The following description of the sample design, including the estimating procedures, and the basis for measuring sampling variation is reproduced with the permission of the Bureau of the Census, from a release, "Concepts and Methods Used in the Current Labor Force Statistics Prepared by the Census Bureau," Current Population Reports, Labor Force Memorandum No. 5, November 8, 1950.

THE SAMPLE DESIGN *

The sample consists of approximately 25,000 households located in 68 sample areas. These sample areas comprise 125 counties and independent cities and the District of Columbia. Each sample area consists of at least one county and may comprise two or three contiguous counties.

Selection of the sample areas. The 3,099 counties of the United States were combined into 2,000 primary sampling units, each of which was defined to be as heterogeneous as possible. A typical primary unit, for example, included both urban and rural residents of both high and low economic levels and provided a broad representation of occupations and industries. Yet, it was sufficiently small in geographic area so that it could be efficiently surveyed without undue travel cost.

These 2,000 primary sampling units were then classified into 68 strata in such a manner that the units within any one class were as much alike as possible. The most important factors in making these groupings were: the degree of urbanization, geographic location, extent of wartime migration, proportion of the labor force engaged in manufacturing, and the type of farming. In some cases stratification by type of manufacturing and by color was also used.

Each of the 12 largest metropolitan areas and the District of Columbia was established as a separate stratum and was included as one of the 68 sample areas. In each of the remaining strata one primary sampling unit was selected in a random manner for inclusion in the sample, the selection having been made in such a manner that the probability of the selection of any one unit was proportionate to its 1940 population.† For example, within a stratum the chance that a primary sampling unit with a population of 50,000 would be selected was twice that for a unit with a population of 25,000. This procedure provides a better representation of the larger units in the sample than does the selection

[13] Morris H. Hansen and William N. Hurwitz, "On the Theory of Sampling from Finite Populations," *Annals of Mathematical Statistics*, 14 (December 1943), pp. 333–62; "A New Sample of the Population," (Washington: Bureau of the Census, September 1944); Edwin D. Goldfield, Joseph Steinberg, and Emmett H. Welch, "The Monthly Report on the Labor Force," *Estadistica*, March 1948, Vol. 6, No. 18, pp. 61–67.

* For a more detailed account of the construction of the sample see U. S. Department of Commerce, Bureau of the Census, *A New Sample of the Population* and *Sampling Methods Applied to Census Work*, by Morris H. Hansen and William N. Hurwitz.

† At the time that the sample was designed, 1940 Census data constituted the most recent benchmark source. As soon as 1950 census results become available, these will be substituted for the 1940 data used in the sample design and estimating procedures.

of units with equal probability (i.e., without regard to size). The selection of a primary unit with probability proportionate to size reduces the sampling variation of the statistics since, in effect, it gives each primary sampling unit a likelihood of selection equal to its influence on the statistics.

Selection of sample households. In determining the approximate number of households to be interviewed in all sample areas, the objective was to obtain a sample that would be of minimum size and yet provide relatively reliable national estimates of the principal labor force statistics.

For each stratum an over-all sampling ratio of 1 in about 1,900 is used. This gives about 25,000 households and units in special dwelling places (such as hotels, institutions, etc.) for the Nation as a whole. The sampling ratio used in each particular sample area depends on the proportion that the sample area population (at the time of the 1940 census) was of the stratum population. Thus, in a sample area which was one tenth of the stratum, the sampling ratio is 1 in 190, which results in a ratio of 1 in about 1,900 for the stratum.

In each area the precise number of households to be interviewed each month is calculated by the application of this sampling ratio rather than through the assignment of a fixed quota. This procedure makes it possible for the sample to reflect any shifts in population. For example, if, on the basis of the 1940 census, a sampling ratio of 1 dwelling unit in every 190 is used in a sample area, the number of households in the sample will be larger than that obtained by a fixed quota in areas where the number of households has increased since the census. In areas where the number of households has declined, the number of sample households will be smaller. In this way, the sample properly reflects the changing distribution of the population during a period of migration, and avoids the distortion which would result from the application of fixed quotas of households or persons based on the population at an earlier date.

In selecting households for inclusion in the sample, an approximate indication was obtained of the number of dwelling units in small geographic areas (blocks, parts of blocks, or similar areas) within each sample area. For the larger urban places information regarding number and location of dwelling units is obtained from large-scale maps which show the general outline of each residential structure. These maps are used for almost every urban place of 25,000 inhabitants or more, as well as for a number of the smaller places. Where such maps are not used, the number of dwelling units in small geographic areas bounded by roads, streams, etc. was determined by field count or by counting from aerial photographs or county highway maps. Thus, each sample area was divided into small geographic areas with well defined boundaries, and for each of which there is a rough indication of the number of dwelling units.

Within each sample area, a designated proportion of the small geographic areas was selected for inclusion in the sample. A list of all dwelling units within each of the selected geographic areas was prepared and clusters containing an average of six contiguous dwelling units were selected for enumeration. In densely populated and urban areas, for which large-scale maps showing the location of each residential structure are available, small areas containing

about six dwelling units were delineated and selected directly, without the necessity of making a list of dwelling units.

In each area the listings of dwelling units in sample segments are brought up to date at frequent intervals, so that each new sample of dwelling units will reflect any new construction or demolition of existing structures.

In August 1949 the coverage of the sample was extended to certain types of special dwelling places (transient hotels, hospitals, and institutions, migratory worker camps, etc.) that previously had been enumerated only occasionally because of the relatively large cost involved. This change was made because residents of special dwelling places have somewhat different personal and economic characteristics from the population as a whole and their exclusion from the sample coverage introduced some bias in the survey results.

ESTIMATING PROCEDURE

The schedules (questionnaire forms) containing the information obtained for each sample household are received in the Washington office by the week after enumeration. The schedules for occupied households for which no interview could be obtained—because of temporary absence, impassable roads due to floods, blizzards, etc., and various other reasons—are replaced by randomly selected schedules for interviewed households of similar race and residence characteristics in the same sample area.

After editing and coding, the raw data are transferred to punch cards, with a separate card for each person enumerated. Estimates could be prepared by tabulating these cards with a fixed weight (the reciprocal of the sampling ratio—approximately 1,900 at present). However, since any sample will tend to vary somewhat in its distribution of basic population characteristcs from that existing in the universe, the accuracy of the labor force statistics derived from the sample is increased by an adjustment of the sample distribution to bring it more closely into agreement with that for the population as a whole. The estimating procedure designed to accomplish this adjustment comprises two steps:

Ratio estimate. The first step adjusts for differences in the distribution by color and by residence (urban, rural-farm and rural-nonfarm). Independently derived distributions for the various residence-color groups are not available on a current basis for the United States. However, a comparison between the 68 sample areas combined and the country as a whole was made from the 1940 census data to obtain adjustment ratios for the residence-color groups. These ratios are applied to the current sample returns, increasing the weight slightly in tabulating for certain residence-color groups and reducing the weight slightly for others. This procedure takes advantage of the correlation which exists between the labor force composition of the primary sampling units currently and their color-residence composition in 1940.

Age-sex adjustment. The second step involves a similar adjustment to take account of differences in the sample distribution of age, sex, and veteran status, for which independent current estimates are available for the entire

population. After completion of the ratio estimate, the sample returns by age, sex, and veteran status are brought into agreement with the independent population estimates by appropriate weighting of the various age-sex groups. These independent estimates are calculated by adjusting the most recent census data to take account of subsequent aging of the population, deaths, and migration between the United States and other countries. A classification of males according to whether they are veterans of World War II is also provided from independent sources.

ADEQUACY OF DATA

Sampling variation. Since the estimates derived from the Current Population Survey are based on a sample, they are subjected to sampling variability. In general, smaller figures and small differences between figures are subject to relatively large sampling variation and should be interpreted with caution. Because mathematical principles are applied in selecting the sample, it is possible to compute with some precision the extent to which the estimates may deviate from a complete census (conducted with the same schedules, instructors and enumerators) as a result of sampling variability.

The interpretation of data in the various labor force reports is made in the light of possible sampling variability. Thus, when two apparently different figures are quoted as "approximately the same" or "not materially different," this citation indicates that the apparent difference may be the result of sample fluctuation.

Estimates of sampling variability are published in Census Bureau reports based on the Current Population Survey. However, because of space and time limitations and the wide variety of data presented, the estimates are often fragmentary or incomplete. The estimates of variability currently presented in the "Monthly Report on the Labor Force" relate only to the numerical levels of the various labor force categories. Percentage figures or distributions are relatively more reliable than the corresponding absolute numbers. Month-to-month changes in levels, particularly, and other changes over periods of time are also subject to less sampling variability than the levels themselves.

Sampling in a census. In the 1950 Population Census the Bureau of the Census obtained certain basic information on every individual. For some subjects, such as income in 1949, residence in 1949 (for identifying migrants), and school attendance, the information was obtained only for a 20-percent sample of the population. By restricting certain questions to a sample, more subjects could be covered at the same cost than if all questions had been asked of everyone. Also, the economies resulting from sampling made possible the use of improved training and supervision in the field that resulted in improvement of the data.

Several alternative designs for sampling were considered. A basic question was whether to use the individual or the household as the unit for sampling, and both methods were pretested. This question was tied in

with the choice of the schedule form to be used—whether to have one schedule for each household or to have a larger schedule with one line for each of 30 individuals, thus including a number of households on one schedule. If the former type of schedule had been adopted, it would have fit well with a plan of sampling each fifth household, while the latter is better adapted to sampling each fifth person.

The larger schedule was finally adopted with one fifth of the lines on each schedule identified as sample lines. The individuals whose names fell on these lines were asked the additional sample questions listed at the bottom of the schedule. If for every schedule the sample lines had been indicated as lines 1, 6, 11, 16, 21, and 26, this would have produced a sample with bias in the direction of too high a proportion of household heads since the prescribed order of listing of individuals in a household is for the head to be first. Bias from this source was avoided by having several forms printed with the sample lines beginning at different lines on each form.

This design of the sample is a systematic design within each enumeration district of the census. Thus, a basic geographic stratification was introduced. The sampling error of this design compared quite closely for virtually all items with those that would have resulted from a simple random sample of individuals. Since lines on the census schedule were sometimes quite properly left blank or contain enumerators' notes, estimates of the total population based on the sample are subject to a small relative sampling error.

This design of sample does not permit certain types of tabulations. For example, if a man who is head of a household falls on a sample line, his wife, who is listed next in order, will not fall on a sample line. Therefore, it is not possible to make a cross tabulation on education of husband by education of wife. If the household had been chosen as the sampling unit and the sample questions asked for each person in the household, this kind of tabulation could have been made.

The use of sampling in the 1950 census just described is only one of the uses that was made of sampling in the 17th Decennial Census of the United States. The population schedule also obtained certain items for only a 3⅓-percent sample, the housing schedule (on the back of the population schedule) obtained information on certain household facilities on a 20 percent sample by asking the first household one pair of questions, the second household a second pair, the third household a still different pair, and so on through the fifth household. Then, the pairs were repeated in order for the next five households and for each succeeding set of five households. A significant part of the 1950 Census of Housing was a sample survey designed to obtain information on financing of residential properties in the United States. The 1950 Census of Agriculture

also used sampling, obtaining data on such items as farm labor and wages, farm expenditures, facilities, and farm machinery on a sample consisting of all very large farms and 20 percent of the remaining farms. Sampling was also used in processing of the 1950 census returns. Preliminary samples were processed to provide early publication of results, both for population and housing and for agriculture. Also, many of the more detailed tabulations of an analytical character are to be made on a sample basis for economy even though the data involved are available on a complete basis.

EXAMPLE OF USE OF AREA SAMPLING WITHIN A STATE

The sampling materials of the Bureau of Agricultural Economics are often used when that bureau carries on cooperative studies with land grant colleges. An example of this type of use is a survey that was made in July–September 1951 to obtain data bearing on various problems in the application of social security to farmers and hired farm workers. This study was a cooperative project of the Departments of Rural Sociology and Agricultural Economics, College of Agriculture, University of Wisconsin, the Bureau of Agricultural Economics, and the Social Security Administration. (A parallel study was also undertaken in cooperation with the Department of Rural Sociology, Connecticut State College of Agriculture.)

In the Wisconsin study the objectives required the use of two contrasting economic areas [14] of the state—one in which farm income was relatively high and in which the practice of hiring regular farm hands was sufficiently common that farm operators would be included who had workers subject to coverage under the 1950 Amendments to the Social Security Act and another area in which the level of farm income was relatively low, in order to ascertain how operators of lower-income farms were facing the problems of providing for their own old age.

The objectives of the study and the budget available were weighed carefully in reaching a decision on the plans for the survey and the sample design. The resulting design was a 2-percent area sample in the open country of each of two economic areas of Wisconsin. One of these was the central part of the eastern dairy belt, an area of highly commercialized dairy farms (designated as state economic area 7 in the delineation released by the Bureau of the Census and the Bureau of Agricultural Economics).[15] The second economic area selected was in the central part

---

[14] Unfortunately, we have to use the word "area" in two senses: (1) to denote the "economic areas" which are groups of counties and (2) to denote process of "area sampling" which involves surveying households living on small areas termed "segments" that contain an average of about six households of farm operators in Wisconsin.

[15] "State Economic Areas of the United States," Census-BAE, No. 15 (Washington: Bureau of the Census and Bureau of Agricultural Economics, August 1950).
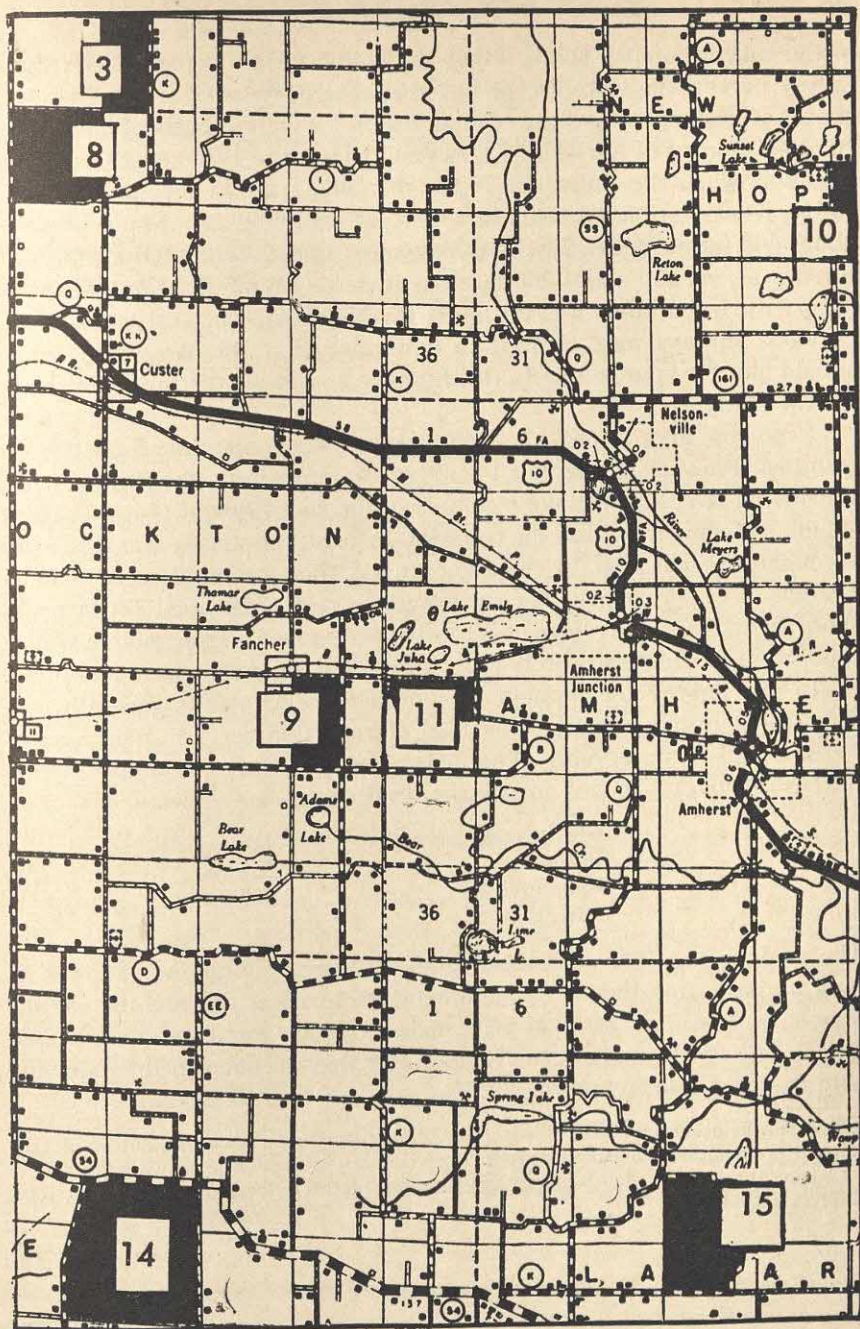
Map 4.   Sample Segments in a Portion of Portage County, Wisconsin.
(Source: General Highway Map, Portage County, Wisconsin, Prepared by
the State Highway Commission of Wisconsin in Cooperation with the United
States Department of Commerce, Bureau of Public Roads. Data Was Ob-
tained from State-wide Highway Planning Survey.)

of the state where the soil is rather sandy and farm income is lower than in any part of the state except in the northern, cutover area. (The area chosen is designated as state economic area 5.) On the basis of data from the 1945 Census of Agriculture (the results from the 1950 census were not available when the study was in the planning stage in the early part of 1951), it was estimated that the 2-percent sampling rate would provide about 400 farms in area 7, with a little more than 100 regularly employed hired workers, and about 200 farms in area 5. (No information or experience with hired labor was obtained in this area). Cost estimates from previous surveys were utilized to determine that the Wisconsin study should aim for approximately this number in view of the funds available for the study.

Economic area 7 contains 7 counties and economic area 5 contains 6 counties. Because of the small number of counties in each, it was not considered advisable to take a sample of the counties and then to take a sample of segments within the counties. Instead, the sample was dispersed throughout each area. In economic area 5 there were 72 segments, an average of 12 per county. The survey enumerators identified 221 farms in these segments,[16] an average of 3.1 farms per segment. In economic area 7, there were 132 segments or 19 per county, and 495 farms were identified.

In any survey the number of completed schedules is generally less than the number of units identified (farms, families, etc.) because of refusals and noninterviews. The latter can be reduced generally by repeated call backs. In this survey the results were as follows:

> Total farm operators..............716
> Interviews.....................657
> Noninterviews.................. 59
> Refusals...................... 29

The map on page 311 of Portage County, Wisconsin, shows the type of map prepared for the survey enumerators to use to locate the sample segments. The segments that were included in the survey were indicated by coloring on the maps actually used, but they are identified by shading here to avoid color reproduction.

---

[16] When the persons planning this study asked the sampling office in the Bureau of Agricultural Economics to prepare maps for a 2-percent sample, that office increased the rate to a 2.5-percent sample to allow for underidentification of farms in the segments, noninterviews, and refusals.
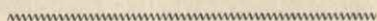
~~~~~~

Tests of Significance of Observed Differences

Transition to statistics of relationship. Again we turn our attention to the procedures of inductive statistics, although we are now in a position to utilize some of the material on application and interpretation of the procedures set forth in the preceding chapters. The reader will recall that in the division of statistical methods according to functions presented in Chapter 1, one of the twofold divisions is into the methods applicable to single distributions and those applicable to two or more distributions considered simultaneously. The methods for treating single distributions are termed "simple" and those for treating two or more distributions simultaneously are termed "complex," or statistics of relationship. In this chapter, the last in the parts devoted to simple statistics, we shall consider what is really a borderline case between single and multiple distributions—the case where we are trying to determine whether two groups of observations represent one or two distributions.

Review of terms. Let us review the use of terms and division of methods as made in this text. Statistical methods analyze numerical data on enumerable nonquantitative characteristics (or traits) and on measurable quantitative characteristics (or variables) for a group or groups of units. A set of enumerations or measures for a group of units with respect to a characteristic, either nonquantitative or quantitative, supplies data on the distribution of that characteristic. With a few exceptions, we have been concerned so far in this text with methods of analyzing, describing, and generalizing applied to one distribution at a time. We have taken up methods especially appropriate for distributions of qualitative attributes in Chapters 7 and 15 and methods appropriate for quantitative variables in Chapters 8, 9, and 16. In Part IV we shall present methods for analyzing simultaneously two or more distributions with various measures of relationship for the various possible combinations of nonquantitative and quantitative distributions.

The simplest case of statistics of relationship. The simplest case of simultaneous consideration of two characteristics, however, is that where

we have information on the distributions of a characteristic from two samples and wish to investigate whether these two distributions differ "significantly" from one another or whether they could have been drawn randomly from the same universe. Since the units are not only measured or enumerated with regard to the characteristic whose distributions are being studied, but are also enumerated with respect to some dichotomous attribute which differentiates between the units in such a way as to cause us to assign some units to one sample and some to another, this is really a case of statistics of relationship. And yet, since the procedures for treating this case are simply extensions of the simple methods for analyzing single distributions, we include their treatment in this part of the text rather than along with the more complex methods.

Meaning of "significant difference." Let us illustrate with one of the most common situations which are treated by the methods of this chapter. Suppose we have measures on the quantitative variable fertility ratio for all counties of the United States and we divide the counties into two groups, rural and urban, which may be considered the categories of a dichotomous attribute, rural and not-rural (or urban and not-urban). Now our problem is to test the significance of the difference between two corresponding summarizing measures of the distributions of fertility ratios in the two groups. Frequently we are interested in testing the significance of the difference between the mean of one distribution and the mean of the other. To test the significance of the difference between two means, we test the null hypothesis that the two means were observed from samples which may have been drawn from the same universe. If the probability of observing two means differing as widely as those actually observed is small enough, we reject the null hypothesis and declare the difference between the two means to be "significant"—that is, it signifies a difference between the means of the universes from which the two sets of observations can be considered random samples. If in the case of counties we establish the fact that the mean fertility ratio of rural counties is significantly higher than that of urban counties, we demonstrate that a positive association exists between the variable fertility and the attribute rurality. Since the conclusions and interpretations of such tests involve the concept of relationships between characteristics, this chapter may serve as an introduction to the statistics of relationship, although utilizing primarily the procedures of simple statistics. And while there is actually an investigation of the association between the dichotomous qualitative characteristic and another characteristic either qualitative or quantitative, the situation is not usually treated with a complete formal analysis of the various aspects of association which are outlined in the next chapter.

Differing sampling situations. As with all applications of sampling theory, the major types of situations are the practical sampling situation

where we have approximately random samples from existent, finite universes and the hypothetical sampling situation where we have two limited universes with data on all their units and have to postulate a hypothetical universe of possibilities. We shall describe the processes of testing the significance of the difference between summarizing measures of both qualitative and quantitative distributions for each of these situations and then continue with special procedures necessary under modified conditions.

TESTS OF HYPOTHESES INVOLVING DISTRIBUTIONS IN TWO SAMPLES FROM EXISTENT LIMITED UNIVERSES

Significance of difference between proportions. To illustrate the test of the significance of the difference between two proportions observed in samples drawn from existent universes, we shall assume that the 117 white tenant farm women for whom data have already been presented are a random sample of all white tenant farm women in a well defined area known as the Tobacco Piedmont. For comparison with them we shall consider a group of 124 white tenant farm women from the states of Georgia, Alabama, Mississippi, and Louisiana, which we shall likewise assume to be a random sample of all white tenant farm women in a well-defined area known as the Deep South. We are assuming further that the sizes of the samples are so small in comparison with the sizes of the universes that we can consider the universes infinite for practical purposes. These two groups will be designated as Piedmont women and Deep South women for brevity.

Of the 117 Piedmont women, 87 or 74.4 percent, and of the 124 Deep South women, 71 or 57.3 percent had never had any occupation other than farming and homemaking. We are interested in testing the significance of the difference between the two percentages, $74.4 - 57.3 = 17.1$, to determine whether there is an association between location and previous occupation among the tenant farm women.

As usual, we set up a null hypothesis that there is no significant difference between the two observed percentages, or, expressed in terms of the universes, that the percentage of those who have never had other occupations among all Piedmont women is equal to the percentage who have never had other occupations among all Deep South women. If such a hypothesis is true, then the difference of 17.1 observed between the two groups must have been caused by variation explainable by chance fluctuation in random sampling. As before, we shall express percentages as proportions before beginning any algebraic manipulations.

Let us adopt the following notation in order to formulate our hypothesis and to compute the necessary measures for testing it:

let N_1 = number of Piedmont women in sample = 117

   N_2 = number of Deep South women in sample = 124

   (A_1) = number of Piedmont women in sample with no previous occupations other than farming and homemaking = 87

   (A_2) = number of Deep South women in sample with no previous occupations other than farming and homemaking = 71

$p_1 = \dfrac{(A_1)}{N_1}$ proportion of Piedmont women in sample with no other previous occupation $= \dfrac{87}{117} = .744$

$p_2 = \dfrac{(A_2)}{N_2}$ proportion of Deep South women in sample with no other previous occupation $= \dfrac{71}{124} = .573$

p_{u_1} = proportion of all Piedmont women with no other previous occupation—unknown

p_{u_2} = proportion of all Deep South women with no other previous occupation—unknown

Expressed in this notation, the null hypothesis which we wish to test is that $p_{u_1} = p_{u_2}$.

Since in the test of a hypothesis every measure computed must be consistent with the hypothesis being tested, we do not need to differentiate between p_{u_1} and p_{u_2} but can let p_u stand for either universe proportion, as we are supposing the two to be equal. Because the estimates of standard errors by which we shall test the hypothesis will be based upon the universe proportion, we must first make an estimate of p_u. The obvious way to estimate p_u is to combine the information from both samples, thus,

$$\hat{p}_u = \frac{(A_1) + (A_2)}{N_1 + N_2} \tag{1}$$

Since $(A_1) = p_1 N_1$ and $(A_2) = p_2 N_2$, the estimate of the common universe proportion is often stated in terms of the sample proportions, thus,

$$\hat{p}_u = \frac{p_1 N_1 + p_2 N_2}{N_1 + N_2} \tag{2}$$

The formula for the standard error of a proportion p_1 observed in a sample of N_1 cases from a universe with a proportion of p_u has been given in Chapter 15,

$$\sigma_{p_1} = \sqrt{\frac{p_u q_u}{N_1}} \tag{3}$$

Therefore, our estimate of the standard deviation of the sampling distribution of p_1 is

$$\hat{\sigma}_{p_1} = \sqrt{\left(\frac{p_1 N_1 + p_2 N_2}{N_1 + N_2}\right)\left(\frac{q_1 N_1 + q_2 N_2}{N_1 + N_2}\right) \times \frac{1}{N_1}} \tag{4}$$

where q, whatever its subscript, is always to equal $1 - p$ with the same subscript. Similarly,

$$\hat{\sigma}_{p_2} = \sqrt{\left(\frac{p_1 N_1 + p_2 N_2}{N_1 + N_2}\right)\left(\frac{q_1 N_1 + q_2 N_2}{N_1 + N_2}\right) \times \frac{1}{N_2}} \tag{5}$$

By a derivation too elaborate to be presented here the standard error of the difference between two independent summarizing measures can be found to be equal to the square root of the sum of the squares of their standard errors. In the case of the two proportions in which we are interested the standard error of their difference is given by the formula,

$$\sigma_{p_1 - p_2} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2} \tag{6}$$

The estimate of the standard error of the difference between p_1 and p_2 is

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\hat{\sigma}_{p_1}^2 + \hat{\sigma}_{p_2}^2} \tag{7}$$

Substituting (4) and (5) in (7) and simplifying, we get

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\left(\frac{p_1 N_1 + p_2 N_2}{N_1 + N_2}\right)\left(\frac{q_1 N_1 + q_2 N_2}{N_1 + N_2}\right)\left(\frac{N_1 + N_2}{N_1 N_2}\right)} \tag{8}$$

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\hat{p}_u \hat{q}_u \left(\frac{N_1 + N_2}{N_1 N_2}\right)} \tag{9}$$

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\hat{p}_u - \hat{p}_u^2 \left(\frac{N_1 + N_2}{N_1 N_2}\right)} \tag{10}[1]$$

$$p_u = \frac{p_1 + p_2}{2} \tag{2A}$$

---

[1] When the numbers of cases in the two samples are equal, that is, when $N_1 = N_2 = N$, these formulas become much simpler. Formula (2) becomes

Formula (8) then becomes

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\left(\frac{p_1 + p_2}{2}\right)\left(\frac{q_1 + q_2}{2}\right)\left(\frac{2}{N}\right)} \tag{8A}$$

Formula (9) becomes

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\frac{2\hat{p}_u \hat{q}_u}{N}} = \sqrt{2}\,\hat{\sigma}_{p_1 \text{ or } 2} \tag{9A}$$

Formula (10) becomes

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\frac{2(\hat{p}_u - p_u^2)}{N}} \tag{10A}$$

In order to test our hypothesis, we shall need to know $\hat{\sigma}_{p_1 - p_2}$. We may get a numerical value for $\hat{\sigma}_{p_1 - p_2}$ by evaluating (8) directly from the data, or by first evaluating (2) to get $\hat{p}_u$ and then evaluating either (9) or (10) to get $\hat{\sigma}_{p_1 - p_2}$. The results will be identical in either case; alternate formulas are offered because, depending upon the form in which data are given, sometimes one is more convenient to use and some times the other. In our example let us evaluate (2) since we are interested in knowing what the estimate of the common universe proportion is. Our estimate of p_u is

$$\hat{p}_u = \frac{(.744)(117) + (.573)(124)}{117 + 124} = .656$$

Since,

$$\hat{q}_u = 1 - \hat{p}_u$$

we find by evaluation,

$$\hat{q}_u = 1 - .656 = .344$$

Now evaluating (9) for an estimate of the standard error of the difference between the two proportions, we have

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{(.656)(.344)\left[\frac{117 + 124}{(117)\,(124)}\right]}$$

$$= .0612$$

It can be theoretically deduced that the sampling distribution of differences between the proportions observed in a sample of N_1 and in a sample of N_2, both random samples drawn from the same universe with the proportion of p_u (that is, the sampling distribution of $p_1 - p_2$), approximates the normal in form with a mean of zero and a standard deviation of $\hat{\sigma}_{p_1 - p_2}$ as already defined. The test then involves finding the probability that a difference in proportions with an absolute value as great as or greater than $|\,p_1 - p_2\,| = |\,.171\,|$ would be observed in such a distribution. As usual, we first express the difference in estimated standard deviation units,

$$\frac{.744 - .573}{.0612} = \frac{.171}{.0612} = 2.79 \text{ standard deviation units}$$

We refer this value to Appendix Table C and find that it corresponds to a probability of .0052. This probability is so low that we reject the hypothesis at the .01 level of significance. Let us organize the above procedures into the five steps we have previously used in testing hypotheses.

1. *Formulation of the hypothesis.* Since we are trying to establish the significance of the difference between the observed proportion .744 of

Piedmont women and the observed proportion .573 of Deep South women who have had no other occupations than farming and homemaking, we shall test the null or opposite hypothesis that these two samples are drawn from universes where the proportions are equal. Since the direction of the difference we have observed is in favor of the Piedmont women, the hypothesis we are interested in establishing is

$$p_{u_1} > p_{u_2}$$

The general null hypothesis which includes all possible alternatives to the above hypothesis is

$$p_{u_1} \leq p_{u_2}$$

The general null hypothesis is too inclusive to be tested. Therefore, we select as the specific null hypothesis to be tested the limiting case of the general null hypothesis,

$$p_{u_1} = p_{u_2}$$

The probability that a difference $p_1 - p_2 > 0$ would be observed under the specific null hypothesis is greater than for any other hypothesis included in the general null hypothesis. Therefore, if we find reason to reject the specific null hypothesis, we can also reject all hypotheses included in the general null hypothesis.

2. *Description of the sampling distribution expected of the difference between the two proportions.* Since in either sample $Np + 9p > 9$, we are justified in using the normal curve to describe approximately the form of the sampling distributions of the individual sample proportions and also the form of the sampling distribution of their difference. We should expect estimates of $p_{u_1} - p_{u_2}$, which for any pair of samples will be $p_1 - p_2$, to have a mean of zero and an estimated standard deviation,

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\hat{p}_u \hat{q}_u \left(\frac{N_1 + N_2}{N_1 N_2} \right)} = .0612$$

3. *Determination of the probability that a difference as unusual as $| .171 |$ would be observed from this distribution.* The observed difference $p_1 - p_2 = .171$ expressed as a deviation from the mean of the sampling distribution in estimated standard deviation units is

$$\frac{(p_1 - p_2) - 0}{\hat{\sigma}_{p_1 - p_2}} = \frac{.171}{.0612} = 2.79$$

From Appendix Table C we find that the probability that a difference as unusual as $| .171 |$ would be observed is .0052.

4. *Rejection of hypothesis.* Since a difference as unusual as that observed would be expected only five out of 1,000 times, we reject the

hypothesis that the proportions in the two universes from which our samples were drawn are equal.

5. *Interpretation of the test.* The rejection of the specific hypothesis $p_{u_1} = p_{u_2}$ implies the rejection of the more inclusive hypothesis that $p_{u_1} \leq p_{u_2}$. The alternative hypothesis which is implicity affirmed is that $p_{u_1} > p_{u_2}$. In the terms of the problem we conclude that a significantly higher proportion of all Piedmont than of all Deep South white tenant farm women have had no previous occupation other than farming and homemaking. If we consider the relationship of the two nonquantitative characteristics, location and previous occupation, we may conclude that among white tenant farm women, these two characteristics are associated —that is, the proportions of individuals having one of the attributes are not the same for the two groups of individuals classified according to the categories of the other attribute. A fuller discussion of the meaning and the description of the several aspects of association will be found in the next part.

Significance of difference between means. In order to focus on the new procedures being introduced, rather than on the variety of problems, we shall test the significance of the difference between the mean number of children ever borne for the same two groups of women, continuing the assumptions as to sampling described above. As to the distribution of children ever borne by the 117 Piedmont women, we have seen in Chapter 8 that the mean number is 6.34, while the corresponding figure for the Deep South women is 5.58. Our problem is to see if the mean number of children ever borne is significantly higher for the Piedmont than for the Deep South group. In beginning to solve the problem, the first step, of course, is to assemble the appropriate data. In this connection it must be remembered that while a hypothesis concerning the difference between proportions can be tested if one knows only the two proportions and the number of cases on which each is based, a hypothesis concerning the difference between two means can be tested only if one knows the two standard deviations of the distributions as well as the means and numbers of cases in the samples.

Letting the subscript 1 refer again to the Piedmont women and the subscript 2 to the Deep South women, we shall use a self-explanatory notation in presenting the formulas for testing the significance of the difference between the two means. These formulas will show in what form it is most convenient to have the data given. The null hypothesis may be formulated thus,[2]

$$M_{u1} - M_{u2} = 0 \text{ or } M_{u1} = M_{u2}$$

---

[2] The more logical notation, $\mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$ is not used because of possible confusion with moment notation.

Next we wish to describe the sampling distribution of $\bar{X}_1 - \bar{X}_2$ or $\hat{M}_{u_1} - \hat{M}_{u_2}$ when $\bar{X}_1$ is based upon N_1 cases with a standard deviation of s_1, and $\bar{X}_2$ is based upon N_2 cases with a standard deviation of s_2. We have seen above that the standard error of the difference between two independent statistics is equal to the square root of the sum of the squares of the standard errors of those statistics. For the case of the difference between means, the standard error may be expressed in formula, thus,

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} \tag{11}$$

But of course since the standard errors of the means are a function of the unknown standard deviations of the respective universes, σ_1 and σ_2, we do not know what values to substitute in the right side of (11). To estimate these two standard errors of sample means, we must first make a decision about how we shall estimate the universe standard deviations. There are two choices here. Since we are testing the hypothesis that the two means come from universes with equal means, we may also add to the hypothesis that they come from universes with the same standard deviation, in which case we shall have to combine the sums of squares from the two samples to find an estimate of the common standard deviation. Unless there is some good reason for doing differently, this is usually the procedure followed and will be the one illustrated here. The other choice is to test the hypothesis that the two means come from universes with equal means but with different standard deviations, each of which must be estimated from its respective sample. The corresponding procedures for this case will be given in a footnote.

The first problem is to estimate the common standard deviation of the two universes. It will be remembered that

$$\hat{\sigma} = \sqrt{\frac{\Sigma x^2}{N - 1}} \tag{12}$$

when the x's are all measured from one mean. If, however, we wish to pool x^2's from two samples to secure an estimate of the common standard deviation, we use the formula,

$$\hat{\sigma}_{1 \text{ or } 2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}} \tag{13}$$

We have seen in Chapter 17 that

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{N}} \tag{14}$$

Substituting (13) in (14) for each sample, we have

$$\hat{\sigma}_{\bar{X}_1} = \frac{\sqrt{\dfrac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}}}{\sqrt{N_1}} \text{ and } \hat{\sigma}_{\bar{X}_2} = \frac{\sqrt{\dfrac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}}}{\sqrt{N_2}} \qquad (15)$$

Substituting the estimates of (15) for the universe values in (11), and simplifying we have

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 N_2}\right)} \qquad (16)\ ^3$$

---

[3] There are two cases commonly met in the testing of differences between means—the one a special case of the above, the other a slightly different case for which we shall give corresponding formulas.

The first case is when $N_1 = N_2 = N$. Because the N's are the same, the formulas become much simpler as may be seen from comparing the four following formulas designated with A's after the number with the corresponding formulas without A's given above.

$$\hat{\sigma}_{1 \text{ or } 2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{2(N-1)}} \qquad (13A)$$

$$\hat{\sigma}_{\bar{X}_{1 \text{ or } 2}} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{2N(N-1)}} \qquad (15A)$$

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N(N-1)}} \qquad (16A)$$

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{2}\,\hat{\sigma}_{\bar{X}_{1 \text{ or } 2}} \qquad (17A)$$

Formula (17A) does not correspond to any of the formulas for the case where $N_1 \neq N_2$ because in that case the estimate of the standard error of the difference between means cannot be reduced to such a simple form. One can see, however, its similarity to formula (9A), the corresponding formula for the standard error of the difference between proportions when $N_1 = N_2$.

The second case is when we have reason to believe that $\sigma_1 \neq \sigma_2$. Then we do not form an estimate of a common standard deviation but use

$$\hat{\sigma}_1 = \sqrt{\frac{\Sigma x_1^2}{N_1 - 1}} \text{ and } \hat{\sigma}_2 = \sqrt{\frac{\Sigma x_2^2}{N_2 - 1}} \qquad (12B)$$

The estimates of the standard errors of the means are

$$\hat{\sigma}_{\bar{X}_1} = \frac{\hat{\sigma}_1}{\sqrt{N_1}} \text{ and } \hat{\sigma}_{\bar{X}_2} = \frac{\hat{\sigma}_2}{\sqrt{N_2}} \qquad (14B)$$

The several formulas for the estimate of the standard error of the difference between the means are

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\hat{\sigma}_{\bar{X}_1}^2 + \hat{\sigma}_{\bar{X}_2}^2} \qquad (11B)$$

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}} \qquad (16B)$$

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{N_2 \hat{\sigma}_1^2 + N_1 \hat{\sigma}_2^2}{N_1 N_2}} \qquad (17B)$$

In actual procedure, one may evaluate (12B), (14B) and (11B); or he may evaluate (12B) and then either (16B) or (17B). The student will find all of these different formulas given in

One may note the resemblance of formula (17) to formula (9), the formula for the standard error of the difference between two proportions. In both cases the estimates of the standard error of the difference between either proportions or means is equal to the expression $\sqrt{\dfrac{N_1 + N_2}{N_1 N_2}}$ multiplied by the estimate of the standard deviation of the universe. (Under certain conditions, $\sqrt{pq}$ may be considered the standard deviation of a distribution of a nonquantitative attribute, when it is treated as a variable with the individuals possessing the attribute given a score of 1, and those not possessing the attribute given a score of 0.)

In our examples we shall choose formula (17) to be evaluated. From page 123 we find that $\Sigma x_1^2 = 1,348.325$ and from similar computations it has been found that $\Sigma x_2^2 = 1,236.1935$. Substituting in formula (17), we have

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{1,348.325 + 1,236.1935}{117 + 124 - 2}\right)\left(\frac{117 + 124}{117 \cdot 124}\right)} = 0.424$$

Since the sampling distribution of differences between means of samples of 117 and 124 has a mean of zero and a standard deviation of 0.424, we can express the observed difference, $\bar{X}_1 - \bar{X}_2 = 6.34 - 5.58 = 0.76$, as a deviation from the mean of zero in terms of standard deviation units,

$$\frac{0.76 - 0.00}{0.424} = 1.79 \text{ standard deviation units}$$

Referring this value to Appendix Table C we have $P = .073$. It depends on what level of certainty we require whether or not we reject the hypothesis when $P = .073$.

Choice of levels of significance; errors of the first and second kinds. In certain fields of research it has become conventional to use the "5-

---

various texts as "the" formula for the standard error of the difference between means. He must realize that all of these are based upon the fundamental formula (11), and that (16) is applicable to the case where $N_1 \neq N_2$, and where σ_1 is believed to be equal to σ_2; that those with A's on their numbers are equivalent and are applicable to the case where $N_1 = N_2 = N$ and where σ_1 is believed to be equal to σ_2; that those with B's on their numbers are all equivalent and are applicable to the case where $N_1 \neq N_2$ and where σ_1 is believed to be unequal to σ_2.

For the other possible case, where $N_1 = N_2 = N$ and where σ_1 is believed to be unequal to σ_2, we can quickly derive from (17B) the only formula that is needed,

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{N}} \tag{17C}$$

In every situation, the student must determine which of the cases his practical problem is, and use the appropriate relations for obtaining his estimate of the standard error of the difference between means.

percent level of significance,"—that is to reject the null hypothesis when $P \leq .05$—in others to use the "one-percent level of significance"—that is to reject the null hypothesis when $P \leq .01$—and still others to use the "one tenth of one-percent level of significance"—that is to reject only when $P \leq .001$. No arbitrary level has been adopted in most of the fields of sociological research, and the choice of level is up to the individual research person. Since it is usually the case that one is trying to establish significance, the selection of the one-percent level or the one tenth of one-percent level is usually regarded as more "conservative" than the choice of the 5-percent level. Yet, errors are made when the arbitrary level is either low or high, and one should examine which type of error is likely to be made most frequently under the different conditions. First, we may explain that a "low level of significance" refers to those levels where the numerical value of P is relatively great, such as $P \leq .1$ or $P \leq .05$; while a "high level of significance" refers to those levels where the numerical value of P is relatively small, such as $P \leq .01$ or $P \leq .001$.

J. Neyman has classified the errors we make in rejecting or not rejecting a hypothesis in a statistical test of a hypothesis into two kinds, which he calls errors of the first kind and errors of the second kind. He defines these errors as follows,

After having applied a test, we may decide to reject the hypothesis H_o, (the null hypothesis), when in fact, though we do not know it, it is actually true. This is called an *error of the first kind*.

After having applied a test, we may decide not to reject H_o (this may be described for short by saying we "accept H_o.") when in fact H_o is wrong, and therefore some alternative hypothesis H' is true. This is called an *error of the second kind*.[4]

Errors of either kind may be committed with either high or low levels of significance, but the lower the level of significance chosen, the more are errors of the first kind likely to be committed; while the higher the level of significance chosen, the more are errors of the second likely to be committed. If, for instance, we choose the 5-percent level of significance (that is to reject hypotheses when $P \leq .05$), we should expect to commit errors of the first kind not more than five out of every 100 times that the hypothesis tested is actually true, for we should expect chance variation to produce such apparently significant deviations with just this frequency if every P were exactly .05. If we choose the one tenth of one-percent level of significance (that is, to reject hypotheses when $P \leq .001$), we should expect to commit an error of the first kind not more than one in 1,000 times that the hypothesis tested is actually true. We cannot state so precisely our maximum expectation of errors of the second kind, but it is

4 J. Neyman, *Lectures and Conferences on Mathematical Statistics* (Washington: The Graduate School of the United States Department of Agriculture, 1938), p. 45.

obvious that if we set up very high levels of significance, differences between samples from universes with different parameters may not be judged to be significant.

The rule usually recommended in choosing levels of significance is to lean over backward in not proving what one wishes to prove. Thus, to protect ourselves from letting our wishes determine our results, when we are anxious to show a significant difference, we should adopt a high level of significance; while when we are anxious to show that no significant difference exists, we should use a low level of significance. As an example, if a southern manufacturer is making a case for regional differentials in wages because of alleged regional differences in cost of living, he should be sure that the difference in cost of living is highly significant (perhaps $P < .01$) to lend weight to his case; whereas if a northern manufacture is making a case for equality in wages because (he claims) there are no significant regional differences in cost of living, he should use a low level of significance (perhaps $P < .05$ or even $P < .1$). It is possible that the manufacturers might do just the reverse of what has been suggested in order to "prove" their cases, but to secure acceptance of results by other scientists, it is always better for a research person to stack the cards *against* himself.

Because there is no uniform convention as to the level of significance to be used in sociological research, it is recommended that for each test of a hypothesis, the actual (or approximate) value of P be presented along with the results in order that the reader may know what level is being used. We do not mean that there should be elaborate interpolation to determine with minute preciseness the value of P. It is usually sufficient to indicate its position as greater than some limiting value and as less than some limiting value, such as $.01 > P > .001$; or if the value of P is very small, it is usually sufficient to indicate that P is less than some limiting value, such as $P < .0001$.

Certain writers have adopted qualifying adjectives to denote certain ranges of P and symbols to indicate these. George W. Snedecor uses the following symbols for ranges of P:[5]

$.05 > P > .01$ —moderately significant—indicated by *
$.01 > P > .001$—highly significant—indicated by **

To these is sometimes added the following,

$.001 > P$ —extremely significant—indicated by ***

Any of these methods of designating the level of significance is satisfactory. And any of them is preferable, in the opinion of the writers, to

---

[5] George W. Snedecor, *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946).

the practice of using only "critical ratios" to denote level of significance. A critical ratio (often designated as C. R.) is the ratio of an observed deviation of a statistic from the mean of its sampling distribution to the standard deviation of its sampling distribution, or what we have been calling a multiple of standard deviation units, with which we enter a probability table. If $2 < $C.R.$ < 3$, we can find from the table of areas under the normal curve that the corresponding level of P is approximately, $.05 > P > .005$. Although results are often still expressed in terms of critical ratios, especially in psychological and educational research, the practice seems less advisable than that of stating the value of P or of indicating its approximate value for two reasons: (1) an older practice was to define the critical ratio as the ratio of the deviation to the *probable* rather than to the *standard* error, and there is likely even now to be confusion as to which definition of critical ratio is intended in a particular report; (2) for the ultimate interpretation one must determine the P corresponding to a given C.R.

To get back to the illustrative example, when $P = .07$, one would probably decide not to reject the null hypothesis although the value .07 is very near .05 level of significance. The value of P is certainly low enough to make us suspicious of the hypothesis, for results as unusual as this would be expected only seven times in 100. It is up to the individual research person to decide whether he is going to regard such a happening as evidence that the hypothesis is untrue. If he does decide to reject the hypothesis he must be sure to state the level of probability which is causing him to reject, since most statisticians do not regard a P of .07 as sufficient evidence for rejection. On the other hand, if he decides not to reject the hypothesis, he must similarly qualify his statement of acceptance. Especially in cases where the value of P is between .1 and .01 it is advisable to state its exact value along with a verdict of a test, for even when we have adopted a .05 level of significance, there is not really much reason for rejecting a hypothesis when P is .045 and accepting it when P is .055. Figure 36 shows graphically the test of the hypothesis that $M_{u_1} = M_{u_2}$. The ratio of the shaded area to all the area under the curve represents the probability of .07.

Again let us summarize this test into the five familiar steps of testing a hypothesis. This particular type of test is so important in sociological research that some repetition is excusable if it fixes the procedures firmly in the reader's mind.

1. *Formulation of the hypothesis to be tested.* To test whether 6.34, the mean number of children ever borne by a group of 117 white tenant farm women in the Piedmont area, is significantly greater than 5.58, the mean number ever borne by a group of 124 white tenant farm women in the Deep South area, we set up the general null hypothesis, $M_{u_1} \leq M_{u_2}$,

and proceed to test its limiting case, $M_{u_1} = M_{u_2}$, under which the probability is greatest for obtaining the observed difference. This hypothesis cannot be tested until it is supplemented with some hypothesis about the standard deviations of the two distributions, and unless there is evidence to the contrary, we usually add to the stated hypothesis a supplementary hypothesis that $\sigma_1 = \sigma_2$. Now it is possible that we might get a verdict of rejection from the test due to the falsity of this second part of the com-

Figure 36. Sampling Distribution of Difference Between Two Means in the Test of the Hypothesis that $M_u - M_2 = 0$.

posite hypothesis, but when such is the case, it would probably have been obvious from inspection of s_1 and s_2 that the supplementary hypothesis should not have been made. For the test, however, we will have to hypothesize either that $\sigma_1 = \sigma_2$ and use a pooled estimate of this value; or that $\sigma_1 \neq \sigma_2$ and use $\acute{\sigma}_1$ based on s_1 as an estimate of σ_1, and $\acute{\sigma}_2$ based on s_2 as an estimate of σ_2. In either case the null hypothesis is composite. In the first case it is $M_{u_1} = M_{u_2}$ and $\sigma_1 = \sigma_2$; in the second case it is $M_{u_1} = M_{u_2}$ and $\sigma_1 = \acute{\sigma}_1$, $\sigma_2 = \acute{\sigma}_2$. We have chosen the former since s_1 is not greatly different from s_2 ($s_1 = 3.45$, $s_2 = 3.16$).

2. *Description of the sampling distribution expected of the difference between the two means of samples.* Under the above hypothesis we should expect the difference between the means of samples of size N_1 and N_2 to be approximately normally distributed with a mean of zero and an estimated standard deviation of

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 N_2}\right)} = 0.424$$

3. *Determination of the probability that a difference between means as unusual as 0.76 would be observed.* The deviation of the observed difference, 0.76 from the mean of the sampling distribution, zero, expressed in standard deviation units is

$$\frac{0.76 - 0.00}{0.424} = 1.79 \text{ standard deviation units}$$

Referring this value to Table C, we find

$$P = .073$$

4. *Nonrejection of the hypothesis.* Since under the hypothesis that $M_{u_1} = M_{u_2}$ (and $\sigma_1 = \sigma_2 = \sigma$) we have found that a difference between sample means as unusual as 0.76, that is, as great in absolute value as 0.76 or greater, would be expected to be observed seven out of 100 times, we cannot reject the hypothesis, although the evidence casts doubts upon it.

5. *Interpretation of the test.* We cannot conclude that the mean number of children borne by all Piedmont white tenant farm women is greater than the mean number borne by all Deep South white tenant farm women, even though the mean number is greater for the observed sample of Piedmont women than for the observed sample of Deep South women. The difference observed is not great enough, in relation to the numbers in the samples and the variability of the individual women in the samples, to be termed a "significant" difference, for differences as great in absolute value would be expected seven out of a 100 times in two samples of the sizes used drawn from the same universe. Let us emphasize, however, that the test has not proved or confirmed the hypothesis that the mean number of children is the same in the two universes. It has simply shown that this is one possible situation which might have led to our getting the results we did. When the verdict of a test of significance is "accept" (nonrejection), no hypothesis is ever proved—in fact, there are no conclusive results in such a case.

Use of the term "significant." In the case of sampling from an existent universe the adjective *significant* applied to some feature of the

description of samples means that the feature is marked or great enough to signify or indicate a corresponding feature in the existent universe. Let us urge that in the reporting of quantitative research the use of the word "significant" be restricted to cases where its technical meaning is the one intended, for otherwise there will be confusion. Although it is true that sometimes differences are so great that an experienced person can immediately judge them to be significant without applying elaborate tests, this is quite a different matter from the practice of inexpert writers who often classify differences as "significant" if their magnitudes are relatively great and as "not significant" if they are relatively small, without any regard for the other variables entering into tests of significance—the numbers of cases in the samples and the dispersions of the distributions. Certainly until one becomes quite expert, he should not label any difference as "significant" without actually making a test of its significance.

When to make tests of significance. We have spoken until now as if research situations always fall into one of two clear-cut types—one type where there is random sampling (or an approximation to it) from an existent universe and the other type where there is complete enumeration of all units of a sort which exist at one time. We have just illustrated the procedures applicable to the first type of situation, and we shall presently illustrate those applicable to the second. We shall interpose between the two, however, a discussion of a type of situation which falls into neither one nor the other of these categories and shall investigate whether or not hypotheses can be meaningfully tested in such situations.

Let us begin to define this type of research situation by first stating what it is not: it is neither a random sample (or a close approximation to one) from a well-defined existent universe, nor is it a complete set of data on all the existing units of a sort. It is, therefore, much more awkward to handle since it is difficult to construct a description of the counterpart of the mathematical model which will serve to tell us what fluctuations to expect from chance variation. And yet, particularly in research where there are comparisons of two groups or "samples," this situation is frequently encountered. A statistician with purist leanings is inclined to discard as valueless for the purpose of statistical inference any data collected in such a fashion, but as long as sociology is in the state of urgently needing to utilize whatever relevant data are offered, it is better at least to consider some possibilities and limitations of the analysis and interpretations of such data.

As an example of the situation to which we are referring, we cite the two groups of white tenant farm women with the assumptions made as to definition of universe and to method of selecting a sample now removed. The true facts of the selection of these two groups are as follows. Because of an interest in southern tenant farm women as a group of high reproduc-

tivity, the writer wished to explore the conditions relating to childbearing and child caring in this group. Because of another focus of interest on a particular group of counties comprising an area of a fair degree of homogeneity in farming economy, the group of 117 Piedmont women were selected from this area, proportionately to the numbers of tenant families in the different counties and within counties proportionately to the numbers in the different tenure classes. The method of selection was by referral from various officials and local people. Although no case could be made for approximating a random method of selection, a case could be made on several points for its representativeness of the limited universe of all white tenant farm married mothers in the group of counties with a recognized bias in the direction of including an undue proportion of those who had borne more children. The sample included, however, two unmarried mothers and one woman (a stepmother) who had actually borne no children of her own. Therefore, since it cannot be strictly called a sample of all women, all married women, or all mothers, it comes very near to being the sort of sample which one writer has castigated as representative of nothing but itself, since one cannot specify precisely the definition of the universe represented. The 124 Deep South mothers were chosen in a similar but even less defensible fashion. Now the particular problem we are interested in here becomes this—in comparisons of summarizing measures of distributions of characteristics among two such groups, such as means and proportions, should tests of significance be made? There is no single criterion on the basis of which the answer can be determined. First, there should be a consideration of the purpose of the inquiry—if tests of significance could be made here validly, would they have meaning in this particular problem? The answer is yes for our example because the purpose of the comparison was to gain some idea of how representative the Piedmont women were of all Southern tenant farm women. Secondly, the matter of the degree of approximation involved in defining the universes from which the two groups might be considered roughly representative samples should be considered. In our example it was the inability to specify precisely the definitions of the universes which led to the decision not to use tests of significance. For instance, the definition of the universe from which the 117 women might be considered an approximately representative sample would have to be defined in some such loose manner as the following—all white tenant farm married mothers and stepmothers, with slightly more than the average number of children, living in a certain group of counties in the Tobacco Piedmont. The definition of the universe from which the 124 Deep South women might be considered an approximately representative sample would have to be even more vaguely defined. When to this vagueness of definitions of the universes represented is added the nonrandom method

of drawing the samples, the situation seems to have departed too far from the ideal for the tests of hypotheses about the universe to have much meaning or validity. Yet, this decision is arbitrary, and another person might have decided differently. If the universes could have been better defined or if the method of selecting the sample had been a closer approximation to the random method, it is probable that tests of significance of differences between the two groups might have seemed advisable.

The chief purpose of elaborating on this example is to illustrate that tests of significance should not be applied blindly regardless of the research situation. Unless their results can be interpreted meaningfully for some universe, real or hypothetical, there seems to be little need for going through the procedures.

On the other hand, by stretching somewhat the concept of the hypothetical universe of possibilities, it is possible to use the concept "significant difference" as a criterion for the stability of a difference. That is, the difference of 17 percent in the proportion of women who have had previous occupations other than farming or homemaking may be evaluated as stable because it is larger than the difference we would expect to arise in random samples from the same universe, and a test of significance may be used to show this. However, the terms "real difference," "true difference," and "significant difference" used in interpreting such a result are likely to be interpreted as though there had been a valid sampling procedure. With the development of the application of statistical analysis to sociological research problems, we may hope that the use of tests of significance under varying situations may be clarified and conventions more firmly established regarding these puzzling matters. At present the best advice that can be offered is for the research person to think through and then to state clearly and exactly what he means in interpreting tests of significance in every situation where there have been gross departures from the ideal case of applying sampling theory. If no clear interpretation can be made of the tests, then they should not be used.

TESTS OF HYPOTHESES INVOLVING DISTRIBUTIONS IN TWO SAMPLES FROM INFINITE HYPOTHETICAL UNIVERSES

Significance of difference between proportions. The procedures used in this situation are identical with those presented for testing the significance of differences between proportions in two samples from existent universes; the interpretation is somewhat different, however. As an illustration we shall consider the difference between the white infant mortality rates for two hypothetical cities for a specified year. An infant mortality rate is the number of deaths of infants under one year per 1,000 live births during the year. It is not exactly the same as the proportion of

babies born alive during the year who die before reaching the age of one. For some of the deaths of babies under one during a year will be of babies born in the preceding year, and some of the babies born during the specified year will live for the rest of that year but die in the succeeding year before reaching the age of one. We shall assume, however, that these two groups will approximately balance each other and that we can treat the infant mortality rate, expressed as a proportion, as if it were the proportion of babies born alive during the specified year who possessed the attribute of dying before reaching the age of one. It is necessary always to examine closely the meaning of a "rate" to see if it can be treated by the methods designed for proportions. Not all "rates" and "ratios" can be so treated. For instance, the common measure of fertility for demographic areas, "ratio of children to women," is not the proportion of women of childbearing age who have borne a child in the past five years. It is rather the mean number who are still alive of the children borne by all women of childbearing age for a five-year period. The fact that when expressed as a decimal fraction rather than as number per 1,000 the "ratio" can take values greater than one shows clearly that the methods for proportions, where p is always less than one, are not applicable. When "rates" or "ratios" are actually condensations of concealed frequency distributions, they must not be treated by the methods for proportions. The fertility "ratio," for instance, is actually the mean of a frequency distribution of the number of children borne during the past five years and still living for all women of childbearing age. The measures which the women have on this distribution will be 0, 1, 2, 3, 4, and even 5 or more in rare cases. Then differences between the means of such distributions cannot be tested unless one knows the standard deviation of the distributions. For the methods of testing differences between proportions to be applicable, one must be able to conceive of the problem as the distribution of some dichotomous nonquantitative attribute among the individuals comprising the base of a "rate" or "ratio." We have just shown that we can do this in the case of infant mortality rates if we make the assumption explained above.

In the example of infant mortality the situation as regards sampling is that we have two existent finite universes, each consisting of observations on all the live births and deaths of infants under one year occurring in a certain area. In testing the significance of the difference between proportions in these two universes, we shall set up and test the hypothesis that both sets of observations might have been observed from one universe of possibilities—that is, that the observed difference is not more unusual than would be expected under conditions of random sampling from the same infinite universe.

Let us test the significance of the difference between the infant mor-

tality rates 30.0 for city A and 23.4 for city B. In thinking of the formulas already given, one immediately realizes that the two rates alone are insufficient data for the test. We must know also the number of cases in the two samples in order to describe the sampling distribution expected of the difference between the rates. The number of live births on which the rate for city A is based is 3,600 and the number for city B is 2,500. Using notation similar to that used above in testing differences between proportions, we can express our necessary data thus,

$$p_1 = .0300 \qquad\qquad N_1 = 3,600$$
$$p_2 = .0234 \qquad\qquad N_2 = 2,500$$

1. *Formulation of the hypothesis to be tested.* Since we are attempting to establish that $p_{u_1} > p_{u_2}$, the general alternative hypothesis is $p_{u_1} \le p_{u_2}$, which has as its limiting case $p_{u_1} = p_{u_2}$, the specific null hypothesis which we shall test. Or expressed in words, the null hypothesis is that a difference of $p_1 - p_2 = .0066$ might be observed between samples of size 3,600 and 2,500 drawn from the same universe.

2. *Description of the sampling distribution of $p_1 - p_2$ expected from samples of 3,600 and 2,500.* First, we estimate the proportion $p_{u_1} = p_{u_2} = p_u$, that is, the common universe proportion from formula (2),

$$p_u = \frac{(.0300)(3,600) + (.0234)(2,500)}{3,600 + 2,500} = .0273$$

Then by formula (9) we estimate the standard error of the difference, $p_1 - p_2$,

$$\sigma_{p_1 - p_2} = \sqrt{(.0273)(.9727)\left(\frac{3,600 + 2,500}{3,600 \times 2,500}\right)} = .00424$$

We then describe the sampling distribution of $p_1 - p_2$ as being approximately normal (since $Np + 9p > 9$) with a mean of zero and a standard deviation of .00424.

3. *Determination of the probability that a difference as unusual as .0066 would be expected from this sampling distribution.* The difference $p_1 - p_2 = .0066$ expressed as a deviation from the mean of the sampling distribution in estimated standard deviation units is

$$\frac{.0066 - 0}{.00424} = 1.57 \text{ standard deviation units}$$

From Appendix Table C we find that the probability that a difference as unusual as .0066 would be observed is .1164.

4. *Nonrejection of the hypothesis.* Since a difference as unusual as that observed would be expected approximately one out of ten times, our

observations do not lead us to reject the hypothesis that $p_{u_1} - p_{u_2} = 0$. Notice that we do not "affirm" the hypothesis tested, for our observations might be, and in fact can be shown to be, consistent with other alternative hypotheses, such as $p_{u_1} - p_{u_2} = .006$.

5. *Interpretation of test.* We have *not* demonstrated that the levels of infant mortality are the same in the infinite universes from which the observation for the two cities may be considered random samples; we have simply failed to show that they are different. Our failure may be due to the fact that there is actually no difference, or it may be due to the fact that the difference is of a magnitude which cannot be detected with samples of the size afforded. The negative aspect of statistical tests of hypotheses must be emphasized again here. So often in research reporting the verdict "no significant difference" is incorrectly used as if it were positive evidence that the universes are the same. Statistical tests, however, are limited to giving two verdicts on hypotheses: (1) the rejection of the hypothesis; or (2) the nonrejection (or "acceptance") of the hypothesis. In the former case we may proceed to say that we "affirm" an alternative hypothesis if it is the only alternative to the rejected one. In the latter case, however, we may not proceed on the basis of the test to reject alternative hypotheses. The interpretation of the test in the illustration is as follows— the observed difference of 6.6 in infant mortality rates of cities A and B is not great enough to afford convincing evidence for believing that the "real" rate of city A is higher than that of city B since so great a difference would be expected from chance variation alone one out of ten times in samples of this size drawn from an infinite universe where the proportions were the same.

Significance of difference between means in two samples from infinite hypothetical universes. Again the procedures for testing differences between the means of limited universes considered as "samples" from hypothetical infinite universes are the same as those for testing differences between means of samples from existent universes. We shall not repeat the procedures for such a test when N is large, and shall defer the explanation of the interpretation of such a test until the next section dealing with small samples.

TESTS OF HYPOTHESES INVOLVING TWO DISTRIBUTIONS WHEN THE NUMBERS OF CASES IN THE SAMPLES ARE SMALL

Testing the significance of the difference between proportions when the number of cases in either or both of the samples is small. The procedures we have presented so far in this chapter are based upon the assumption that the form of the sampling distribution of the differences between proportions of two samples is normal or so close an approximation to

normal that the table of areas under the normal curve can be used to determine the probability that a difference as unusual as the observed would be expected. We have seen in Chapter 15 that the sampling distribution of a proportion in samples of N cases cannot be considered continuous or of normal form when N is very small or when p ($p < q$) is very small. The sampling distribution of the differences between proportions in samples of N_1 and N_2 cases will not depart farther from normality than the sampling distribution of the one of the proportions which departs the farthest. Therefore, we can test the sampling distribution of the proportion in the sample containing the fewest cases by the rule-of-thumb procedure suggested in Chapter 15, and if we find that it is safe to use the normal distribution to describe this sampling distribution, it will be safe to use the normal distribution to describe the sampling distribution of the difference between proportions observed in this sample and in another sample with a greater number of cases. The p substituted in the required relation, $Np + 9p > 9$, should be the estimated common universe proportion, which we have designated as p_u (when $p_u < q_u$), and the N substituted in the relation should be N_1 (when $N_1 < N_2$).

Let us apply this test to the example already given of testing the significance of the difference between the proportions of women who have had no occupations other than farming and homemaking in the groups of Piedmont and Deep South women. It will be necessary to use for p_u the value previously designated as q_u, since it is the smaller of the two, and to use $N_1 = 117$ for N since this is smaller than $N_2 = 124$. Then

$$N_1 p_u + 9 p_u = 117(.344) + 9(.344) = 42.1 > 9$$

where

$$p_u > q_u \qquad \text{and } N_1 > N_2$$

Since 42.1 is greater than 9, we see that the normal distribution can be safely used to describe the sampling distribution of $p_1 - p_2$, as we did in testing the significance of the difference above.

An alternative method of treating this case by the use of the chi square distribution will be given in Chapter 21. The method, however, is applicable for about the same range of values of Np as the above procedures.

Testing the significance of the difference between means when the number of cases in either or both samples is small. No matter which of the several formulas given is used for obtaining the estimate of the standard deviation of the sampling distribution of the difference between means of samples of size N_1 and N_2, somewhere in the computations there will be a division by the expression $N_1 + N_2 - 2$. This expression represents the

number of degrees of freedom on which the estimate is based, in this case the total number of observations less two. The assumption that the sampling distribution of the differences between means approaches normality closely enough for us to use the table of areas under the normal curve is justified only if $N_1 + N_2 - 2$ is large.

The more accurate description of the form of the sampling distribution of a statistic (such as the mean or the difference between two means) divided by the *estimated* standard error of the statistic, is Student's t distribution, which as we have seen is slightly different for each different value of n, the number of degrees of freedom. Theoretically, in the case of testing hypotheses about any parameter when the standard deviation of the sampling distribution of its corresponding statistic has to be estimated, one should always use the t distribution rather than the normal distribution to describe the sampling distribution of the statistic. As in the case of testing hypotheses about the mean of a single distribution, however, the use of tables based upon Student's distribution give results practically the same as those based upon the normal if $n = N_1 + N_2 - 2 > 100$ in the case of testing hypotheses about universe parameters corresponding to the difference between means of samples of N_1 and N_2. The difference in results is not of much importance unless $n = N_1 + N_2 - 2 \leq 30$.

Example.[6]   In a study of the Aid to Dependent Children program in 1950 information was obtained on a sample of all Aid to Dependent Children cases (hereafter abbreviated as ADC cases) in the United States. One item of information obtained on each case studied was the nature of the crisis that eventually caused the family to have to appeal to the ADC program for financial aid. The length of time between the occurrence of this crisis and the time when aid was actually received (known as the crisis period) was ascertained for each family studied (note that this is not the length of time between applying for aid and receiving it). The 20 sample cases from Denver County, Colorado, were divided into two groups. Group 1 was made up of those cases in which the crisis was due to the father, that is, his death, desertion, or becoming incapacitated. Group 2 was composed of all other cases. The length of the crisis period in months was tabulated for each case in each of these groups. Data on this item are shown below:

Group 1	Group 2
$N_1 = 11$	$N_2 = 9$
$\bar{X}_1 = 30.6$	$\bar{X}_2 = 7.00$
$s_1 = 27.01$	$s_2 = 5.395$
$\Sigma x^2 = 8,026.54$	$\Sigma x^2 = 262.00$

---

[6] The data for this example are taken from a study of the Aid to Dependent Children program made for the American Public Welfare Association by the Institute for Research in Social Science, University of North Carolina at Chapel Hill, during 1950 and 1951 by Gordon W. Blackwell and Raymond F. Gould.

For these two groups the mean length of the crisis period is much greater for group 1 where the crisis was related to the father. Let us test the significance of the difference between these two means in the customary five steps.

1. *Formulation of the hypothesis.* The observed means indicate that $M_{u_1} > M_{u_2}$. To establish this hypothesis we set up the general null hypothesis which covers all alternatives $M_{u_1} \leq M_{u_2}$ and choose its limiting case $M_{u_1} = M_{u_2}$ as the specific null hypothesis to be tested.

2. *Description of the sampling distribution of the difference between means.* Under the null hypothesis the mean of the sampling distribution of the difference between the means of two samples is zero. Since

$$n = N_1 + N_2 - 2 = 18$$

we shall use Student's distribution to describe the form of the sampling distribution rather than assuming that it is normal. The estimated standard error of the sampling distribution is found by evaluating formula (16), thus,

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(\frac{8,026.54 + 262.0}{9 + 11 - 2}\right)\left(\frac{9 + 11}{9 \cdot 11}\right)} = 9.645$$

3. *Determination of the probability that a difference as unusual as 23.6 months would be observed in this sampling distribution.* The ratio of the difference of the observed means to the standard error of their difference is

$$t = \frac{30.6 - 7.0}{9.64} = 2.45$$

The estimate of the standard error of the difference is based on $N_1' + N_2 - 2 = 18$ degrees of freedom. In Appendix Table D we find that the value of t required for a probability of .05 with 18 degrees of freedom is 2.101, and the value required for a probability of .02 is 2.552. We can without interpolation indicate the value of our P, thus,

$$.05 > P > .02$$

4. *Rejection of the hypothesis.* If we use the .05 level of significance, we reject the hypothesis that $M_{u_1} - M_{u_2} = 0$.

5. *Interpretation of the test.* The sampling situation here was designed to be a close approximation to a random sample. For this reason we can conclude that in Denver County, Colorado, the crisis period was longer for ADC families of type 1 than for ADC families of type 2, and we feel relatively safe in this conclusion in spite of the fact that we only had eleven families of one type and nine families of the other. We must be careful, however, not to generalize from this sample to other areas. We know that this situation is true in Denver County, but we cannot say

that it is true in the county next to it. By testing this same hypothesis with other samples from other areas, we might arrive at the conclusion that this situation is true only in urban areas, or we might find it to be true only in certain regions of the United States, or we may not find any other areas where it is true.

In dealing with such small numbers of cases, it is necessary to check statistical results with nonquantitative information continually. Are our results consistent with what one would anticipate? It seems reasonable that in cases where the crisis was the desertion of the father there might be some hesitancy on the part of the mother to apply for aid, though more information and more detailed analysis would be needed to demonstrate that this actually was the situation.

CRITERIA FOR APPLICABILITY OF THE PROCEDURES PRESENTED IN THIS CHAPTER

As we have stated, tests of significance of differences in summarizing measures of two distributions of the same characteristic should not be made blindly whenever a comparison is possible. One of the two principal criteria has already been set forth; the other, since it involves some principles treated in the statistics of relationship, has not yet been explained fully. We shall state explicitly these two criteria here, giving special attention to the very important one not discussed previously. It is suggested that the student just beginning to learn to use tests of significance check carefully each situation where he is planning to use such tests on these criteria before making the tests.

Criterion of relevance. The sampling situation must be such that generalizing to the universes, existent or hypothetical, from which the two groups of observations compared may be considered random samples has a meaningful interpretation relevant to the purpose of the inquiry.

Criterion of independence. By using the concept of a hypothetical universe, tests of hypotheses are often utilized in research situations where there is no actual sampling but instead a complete enumeration of two existent universes. In this situation, however, as well as in the situation where there is actual sampling from an existent universe, one extremely important criterion must be met for the above procedures, based upon random sampling theory, to be applicable—the criterion of independence of the sets of observations being compared. It will be remembered that the description of random sampling includes the provision that the inclusion in the sample of one unit must in no way affect the probability of any other unit's being included in the sample. Furthermore, the methods presented in this chapter for testing hypotheses about distributions observed in two samples are based on the assumption that the inclusion

of a unit in one sample in no way affects the probability of the inclusion of units in the other sample. Therefore, one cannot test hypotheses relating to two sets of observations by these methods if the units included in one set of observations are dependent in any way upon the units included in the other set. The concepts of association and independence cannot be fully treated until Part IV, but we can at least illustrate here the sort of situation where the sets of observations are not independent and the procedures of this chapter cannot be applied.

Let us consider a study of N pairs of married couples, in which we are interested in testing the significance of the difference between the mean age of the husbands and the mean age of the wives. Since for every man included in the group of husbands, the wife is included in the group of wives, the two sets of observations of ages, one of the husbands' the other of the wives', are not independent. Sometimes the lack of independence is evident from the method of selection of samples as in this case; another way of demonstrating the lack of independence would be to compute the coefficient of correlation between the ages of husbands and wives. There are methods for treating such cases which take into account the lack of independence of the two sets of observations, but they will not be presented until measures of relationship have been treated. Other situations where the sets of observations are not independent are found when there is an overlapping of the sets of observations. This is often found in two sets of observations made on the same units at different times. There will be some discussion later on methods for treating some of the situations where the summarizing measures are not based on independent samples, although methods have not yet been developed for certain situations of this type. The point we wish to make just here, however, is that the methods of this chapter must not be used unless the two sets of observations are independent.

SUGGESTED READINGS

Fisher, R. A., *Statistical Methods for Research Workers*, 10th ed. (London: Oliver and Boyd, 1948), Chap. 5.

Peatman, John Gray, *Descriptive and Sampling Statistics* (New York: Harper, 1947), Chap. 14.

Smith, John H., *Tests of Significance, What They Mean, and How to Use Them* (Chicago: University of Chicago Press, 1939).

Snedecor, George W., *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), Chaps. 2 and 4.

Treloar, Alan E., *Elements of Statistical Reasoning* (New York: Wiley, 1939), pp. 138–151, 200–209.

PART IV

Statistics of Relationship

Introduction to Statistics of Relationship

PURPOSE AND LIMITATIONS OF STATISTICS OF RELATIONSHIP

Simple statistics. The statistical methods presented in Parts II and III have been concerned with describing single distributions of non-quantitative or quantitative characteristics among groups of varying units for the observed groups of units, for existent finite universes, or for infinite hypothetical universes. Sometimes the summarizing measures comprising the descriptions have been exact, when there are data on all units; sometimes they have been estimates, where there are data on only a fraction of the units. In all cases (with the exception of parts of Chapter 19), however, the summarizing measures have referred to one distribution of a single characteristic, and the methods for making the descriptions have hence been termed "simple." Such methods are limited to the function of describing the quantitative features of distributions of characteristics considered one at a time.

The problem of relationship. The aim of scientific research in any field goes beyond describing the incidence of distributions of phenomena with each class of phenomena (characteristic) considered separately, although this is a necessary first step. In order to achieve the ultimate goals of science—prediction and mastery—research must not stop with describing the manifestations of one characteristic alone as it is distributed among varying units, but must also describe the conditions under which its different degrees of incidence occur. Since we are using the term "characteristic" in a broad sense, the conditions may also be referred to as characteristics. Therefore, the general problem of science becomes the discovery, analysis, and verification of relationships between two or more characteristics. Insofar as one or more of the characteristics are regarded as "conditions" for the occurrence of one or more other characteristics, the interrelationships between the characteristics investigated by the scientist are what the layman thinks of as "cause and effect" relationships.

We hesitate to use the words "cause and effect," not because they do not have the same practical meaning for the scientist as for the layman but because they have for long had metaphysical implications that the explanation of cause is to be found in some inherent quality of the phenomena beyond the realm of scientific observation and inquiry. The scientist is not concerned with such considerations. He considers that he has "explained" phenomena when he has discovered the conditions under which the phenomena occur. To achieve the explanation he must investigate all the relevant conditions (other characteristics) which are associated with the phenomena and determine which ones are actually associated and which ones are only apparently associated. By "actually associated" we mean that two characteristics are *related* in the everyday sense of the word, and by "apparently associated" we mean that two characteristics which are *not* related in the everyday sense of the word have been observed to occur together. An analysis of relationship results in an "explanation" of a series of phenomena when the relationships identified can be used to predict future levels of occurrence of the phenomena, given information on the actually associated phenomena. We shall differentiate between the terms "association" and "relationship" on this basis, using relationship to mean an "actual association" with predictive validity as a criterion and using "association" to mean any observed simultaneous occurrence between two or more characteristics regardless of the predictive value of measures describing the association.

The role of statistics in the scientific investigation of phenomena. Statistics alone does not supply the methods for all the steps in a scientific problem of relationship. Let us examine the potentialities and limitations of statistics in this area.

First, in keeping with our definition of statistics, we observe that it is useful only when the problem of relationship is between characteristics for which we have observations on a *plurality* of units. The methods of statistics are not fitted for analyzing and describing either single characteristics or relationships between two or more characteristics manifested by a single unit; they are designed and may be used only for the analysis and description of the incidence of single characteristics or of relationships of two or more characteristics as they are distributed among a number of units. Thus, if the high income of a particular father makes it possible for him to send his son through college, we cannot use statistical methods to study the relationship between the father's income and the grades of education completed by the son in that particular family. If, however, in a survey of many out-of-school youths we obtain data on the annual income of the father of each youth and on the number of grades of school he has completed, we may use the methods of correlation to investigate the relationship between father's income and youth's education as these

characteristics are distributed in the group of youths surveyed. The sort of relationship investigated by statistics is always an *average* relationship existing between two or more characteristics distributed among a series of units. Consequently, as even the student acquainted with only simple statistics would anticipate, the predictions made from statistical analysis of relationships will be in terms of group averages with an expected range of error, which can often be estimated.

Secondly, statistics is limited to investigating relationships between characteristics for which measuring devices have been developed. When there are known to be relevant factors for which measuring devices have not been developed, then their bearing on any problem of relationship is not amenable to statistical treatment. Consequently, the utility of statistics in the analysis of relationships is limited to the "explanation," often times partial, of those measurable or enumerable characteristics, for which the identification and verification of relationships with other enumerable or measurable characteristics comprise a substantial part of their "explanation."

Thirdly, in the broader problem of relationship, statistics proper is limited to the function of analyzing the several aspects of association between two or more characteristics—the existence, direction, degree, and nature (in a technical sense to be explained shortly) of the association. While certain statistical techniques can be utilized in designing a study for the analysis of relationship, statistics proper can neither select the factors which are to be considered nor determine when all relevant conditions have been taken into account. Therefore, the summarizing measures of association provided by statistics are not in themselves conclusive evidence of relationship until they are supplemented by nonstatistical evidence that all relevant factors have been considered in the analysis. The co-occurrence or statistical association of phenomena is by no means a sufficient condition to justify the interpretation of relationship between phenomena. Yet, it is one of the necessary conditions, and it is the one on which statistics can offer evidence. This point is so important as to deserve elaboration. The body of methods termed statistics of relationship contains procedures for investigating and describing the several aspects of association between two or more series of phenomena by various summarizing measures of association. The summarizing measures constitute the contribution of statistics in the larger problem of investigating and establishing relationships. The maximum information that they can offer is that two classes of phenomena always vary together; usually they offer evidence of a less than perfect degree of association. They are, therefore, an important and one of the most highly developed set of methods for attacking the problem of relationship, but those who are most familiar with the methods are the first to admit that they contribute to only one

phase of the problem—the establishing of one necessary condition of relationship—association between two classes of phenomena. There are other conditions also necessary before relationship is demonstrated and these must be established by nonstatistical methods.

If all those who use statistics, and their critics as well, comprehended thoroughly both the utility and the limitations of statistics of relationship in the problem of discovering and establishing relationships, there would be no grounds left for disagreement over the statistical part of the interpretation of correlation coefficients and other summarizing measures of association. Statistical methods have finished their part of the task when they have provided a statistical description of association, and statistics proper cannot be held responsible for the further use of the data it supplies for the larger problem of relationship.

We are not proposing that the research worker stop with these measures. He should, of course, go on to the major problem of relationship which he is attacking, using these measures as one type of data. Unfortunately, however, statistics cannot offer him all the methods necessary for solving his problem. A noted statistician has said that when one stops with a correlation coefficient, it is an admission of failure. Stopping with the description of association without determining its relevance in the larger problem of relationship is one of the reasons that studies utilizing statistical analysis have not led to more fruitful results; another reason is the mistake of some writers in believing that statistical measures of association invariably mean actual relationships and in so interpreting their results. It is easier to avoid the second reason for failure than the first, however. So many writers on the use of statistics have emphasized (almost to the exclusion of any constructive suggestions) the things *not* to do in statistics, that only a person completely unread in the literature available on the application of statistical methods to sociological problems would make the mistake of automatically interpreting a correlation coefficient to mean conclusive evidence of relationship. On the other hand, one has to search far and wide to find constructive suggestions, or even excellent examples, of how to proceed with the problem of relationship beyond the correlation coefficient. There is urgently needed an exposition of the logic involved with illustrations in the sociological field. And yet, in any one particular problem, "going beyond the correlation coefficient" requires not only a knowledge of the logic and principles of the scientific treatment of relationship, but also an intimate and, if possible, exhaustive knowledge of the subject matter being treated.

The paragraphs above have attempted to give a somewhat cursory treatment of the purpose, utility, and limitations of statistics of relationship in order to show the reasons for studying the methods in this part.

We shall now look at the content of the methods and at the way they are organized in this text.

METHODS OF STATISTICS OF RELATIONSHIP

Aspects of association investigated.   Statistics of relationship deals with the analysis and description of several aspects of the association between the distributions of two or more characteristics among the same group of units. We use "association" here instead of "relationship," for at this stage of statistical analysis in a problem of relationship we cannot usually know whether an observed association will prove to be indicative of a relationship, as both terms were differentiated earlier with relationship specifying an association whose predictive validity has been established. The analysis of association between two or more distributions of characteristics is usually divided into four parts, each concerned with elucidating a particular aspect of the association. The aspects with the particular questions to be answered about each are summarized below.

ASPECTS OF ASSOCIATION INVESTIGATED BY STATISTICS

1. *Existence.* Is there any association between the characteristics?
2. *Direction.* Are the characteristics positively or negatively associated?
3. *Degree.* How close is the association?
4. *Nature.* (For nonquantitative characteristics) What levels of incidence of the categories of one characteristic are associated with the categories of another characteristic?

(For quantitative characteristics) What changes in the measures of one characteristic are associated with unit changes in the measures of another characteristic at the varying levels of the second characteristic?

Bodies of method included in statistics of relationship.   Various summarizing measures have been devised to answer these questions about the association between the distributions of two or more characteristics. Different types of summarizing measures are used for different combinations of the two types of characteristics, nonquantitative and quantitative. The procedures for computing these measures are usually grouped into bodies of method applicable to the various combinations of types of characteristics studied and are identified by some one term which indicates the outstanding feature of that body of method. The grouping of methods can be summarized as shown on page 348.

In each of these bodies of method there are procedures for investigating the four aspects of association listed above: existence, direction, degree, and nature of association. Furthermore, in each body of method

Combination of Types of Characteristics	Body of Method for Investigating Association
1. Two or more nonquantitative characteristics	1. Contingency
2. One quantitative characteristic and one or more nonquantitative characteristics	2. Analysis of variance
3. Two quantitative characteristics	3. Total correlation and regression
4. Two or more quantitative characteristics and one or more nonquantitative characteristics	4. Analysis of covariance
5. Three or more quantitative characteristics	5. Partial and multiple correlation, regression, and factor analysis

there are procedures for describing these various aspects of association in the sample studied and for estimating their descriptions in the universe from which the sample has been drawn. That is, both the descriptive and the generalizing functions of statistics may be utilized in the investigation of association. The plan of presenting statistics of relationship in this text will be to study the several bodies of method as listed, and under each body the several aspects of association as listed, first the procedures for computing the summarizing measures which describe that aspect for the sample, then the procedures for making estimates of corresponding summarizing measures for the universe with their appropriate measures of error and precision, and finally, the procedures for testing hypotheses about the universe parameters.

TYPES OF CHARACTERISTICS FOR WHICH THE METHODS OF STATISTICS OF RELATIONSHIP ARE APPROPRIATE

The above outline of the bodies of method included in statistics of relationship designates the types of characteristics for which they are appropriate only as "nonquantitative" or "quantitative." We have seen in Chapter 6, however, that there are gradations of types of characteristics in the nonquantitative-quantitative dimension, and we need to specify more precisely the situations for which the various bodies of method are appropriate. This will be done with reference to the classification of characteristics given on page 66.

The associations between characteristics of type I are usually analyzed and described by the methods of contingency to be presented in Chapter 21. It is possible also to use techniques giving tetrachoric coefficients of correlation, if one assumes that the categories of the nonquantitative characteristics are really twofold class intervals of normally distributed

variables. Or it is possible to compute a modified Pearsonian coefficient of correlation under certain assumptions. However, the simpler methods described under "contingency" are most frequently employed.

The associations between characteristics of type III are usually analyzed and described by the methods of correlation to be presented in Chapters 23 and 25. When the distributions of two characteristics are cross-tabulated into class intervals, however, it is possible to reduce the data to the form of a fourfold or a manifold contingency table and to use the methods of contingency. Information is sacrificed by use of the simpler methods, however, and furthermore, the value of the coefficient of association so obtained depends on where the division into two or more parts is made along the scale of the variable. Therefore, although other methods are possible, the data on association between characteristics of type III are most frequently analyzed by methods of correlation.

It is the intermediate group of type II characteristics which offers more choice in selection of methods. Modifications of the methods of either contingency or correlation are applicable to them. Although various treatments are possible, the following practice is recommended. If the data on the distribution of a characteristic designated as type II A are given as frequencies in categories designated by nonquantitative descriptions, which nevertheless are definitely ordered, such as "good health," "fair health," "bad health," the methods of contingency for manifold classifications are usually appropriate. If for a characteristic designated as type II B the data are given by ranks assigned to the unit, denoting the relative degree of the characteristic possessed, the methods of rank order correlation are appropriate. If for a characteristic designated as type II C the data are given as scale values such as one obtains from an attitude scale, the regular methods of correlation are advised, although there are those who criticize the treating of scale intervals as equal when equality has not been demonstrated.

When one of the characteristics whose association is being investigated is of one type and the other of another type, the situation becomes more complex, and when there are more than two characteristics of different types being studied simultaneously, it becomes even more complex. We usually decide to treat each characteristic as nonquantitative or quantitative—that is, as a type I or a type III characteristic—and choose analysis of variance or covariance if both types of characteristics are involved.[1]

---

[1] Yet see Milton Friedman, "The Use of Ranks in Analysis of Variance to Avoid the Assumption of Normality," *Journal of the American Statistical Association*, 32 (December 1937), pp. 675–701.

TYPES OF VARYING UNITS MANIFESTING ASSOCIATION
OF CHARACTERISTICS

Characteristics are abstractions; we can measure them to secure data for analysis only as they are manifested or possessed in varying degrees by real observable units. The characteristic "fertility," for instance, cannot be studied statistically until we specify the type of unit which we shall regard as manifesting or possessing the characteristic "fertility." This particular characteristic has meaning for both of the two major types of units used in sociological studies—an individual human being and a population unit composed of many human beings. For the first type of unit women within certain age limits might be chosen and the number of children ever borne by each be considered as the degree of the quantitative characteristic possessed by each varying unit. For the second type of unit the populations of a series of counties might be chosen and the ratio of children under five to women of childbearing age be considered as the degree of the quantitative characteristic possessed by each varying unit. Most of the methods and theory of statistics of relationship, and especially of correlation, were developed in relation to the first type of unit —for the description of the association of characteristics distributed among human beings. Increasingly, however, the methods have been adopted by those interested in studying variation of group characteristics. We shall proceed to examine the differences of situation involved in problems of this second type, particularly in the case of demographic units.

Differences in measures of characteristics for different types of units. When a single individual is considered as the varying unit, the problem of measurement of degree of incidence of characteristics is relatively more direct. Consider, for example, the process of measuring for an individual height, weight, age, I. Q. (as defined by some standardized test), annual income, or even less tangible characteristics, such as neuroticism or unfavorableness toward the Negro. In each case there is a technique devised for measuring directly some property or behavior of the person, regardless of whether the measure obtained is considered as a direct or indirect measure of the characteristic being studied. When the population of an area such as a county is considered as a varying unit, the process of getting a measure on some characteristic is less likely to correspond to our everyday notions of measuring. The measures are most frequently ratios, percentages, or rates computed from data secured through enumeration of certain items of information about each person in the area once in ten years by the census and through the continuous recording of data for the area as a whole by those agencies dealing with various aspects of the functioning of a county—public health, welfare, education, government,

etc. Since for comparison with other counties, relative rather than absolute measures of the incidence of a characteristic are desired, any single count of individuals possessing an attribute, or of events, will usually be reduced to a per capita basis by forming a ratio, percentage, or rate.

There is confusion and disagreement over the meaning of the use of certain measures for demographic units. For instance, a death rate of 14.7 per 1,000 population is often called a measure of the "force of mortality" on a demographic unit for a given year. It does not mean, however, that each person in the area had an equal probability of $\frac{14.7}{1,000} = .0147$ of dying within that year. We know that individuals under one and over 60 are much more likely to die than those in between these two ages. Yet, the rate 14.7 does measure the effect of mortality on the population of that county as a whole, and it measures a characteristic which has no exact parallel if we were concerned with individuals as varying units. Many demographic characteristics, such as age composition, sex composition, annual rate of natural increase, percentage of net decennial change through migration, suicide rates, etc., have been defined and dealt with as characteristics possessed in varying degrees by demographic units or populations as wholes. It is this type of measure for this type of unit with which we shall be especially concerned because when sociologists need to be concerned with analyzing associations of characteristics pertaining to individuals, they have the advantage of being able to find much more writing and many more examples in the literature of related fields than they can find for demographic characteristics of populations.

Distributions of quantitative demographic characteristics. To carry over the methods developed for studying relationships between characteristics distributed among individual human beings to problems of demographic characteristics distributed among populations requires close examination of conditions of applicability in the latter situation. First, we may consider the form of the second type of distributions since certain methods to be treated are based upon the assumption that the distributions whose associations are being analyzed are normal. Most of the distributions of demographic characteristics are not distributed normally among demographic units; they are likely to be skewed to the right. Any set of measures on a characteristic which has a definite lower limit of incidence, such as zero, within several standard deviation units of the mean is likely to have the upper limit of its range farther from the mean than its lower limit. Many sorts of per capita measures, such as percentage of illiterates, are of this type. The general practice is to go ahead and treat such distributions as if they were normally distributed, if they are I-type curves and not too extremely skewed. In correlation analysis, how-

ever, one or two very extreme measures may have more weight than all of the others in a distribution and give misleading coefficients. Therefore, the effect of extreme cases should be carefully examined in making a decision as to whether or not correlation methods are valid in a particular problem.

Measures of the characteristic of change of a characteristic, which can take either positive or negative values, are more likely to approach a normal distribution, for there are no limiting values on either side of the mean. Such a characteristic is "percentage net change due to migration, 1940–1950." It will often be found that while the distribution of the levels of incidence of a characteristic may be too far from normal for correlation methods to be applicable, changes in levels of the characteristic from one period to another may approximate closely a normal distribution. Since the sociologist is interested perhaps more in changes and their association with other factors than in absolute levels, this type of problem is one type where correlation is frequently used.

Lack of independence of observations. Not only correlation analysis, but also almost all statistical theory is based upon the assumption that any series of N measures made on N units are *independent* observations. Never is this assumption valid in problems involving measures of all the demographic units of one nation. The contiguity of adjacent areal units is the most obvious evidence of lack of independence between the levels of incidence of demographic characteristics. An analogous problem has been faced by economists and others in the lack of independence in measures of a characteristic for successive years on the same unit. There have been developed methods for estimating the number of *equivalent* independent observations represented by N observations made in successive time intervals.[2] To the writers' knowledge no such solution has been offered for the problem of lack of independence in observations on units contiguous in space rather than in time.

Arbitrary nature of varying units. Ward, township, county, and state lines, which delimit the demographic units we have to work with, are often the result of political accident. A county is not necessarily a homogeneous unit with respect to demographic characteristics, nor is it so much a demographic entity as it is an arbitrarily delimited segment of population which we measure as a unit. Census and other data are gathered and tabulated by political divisions, and we have to accept the information in the lumps in which it comes. Efforts to make the census units correspond to "natural" areas have been made in the delineation of

---

[2] L. R. Hafstad, "On the Bartels Technique for Time-Series Analysis, and Its Relation to the Analysis of Variance," *Journal of the American Statistical Association*, 35 (June 1940), pp. 347–361.

census tracts in selected cities and in the delineation of "economic areas" [3] for the whole United States, for which certain data have been tabulated for the 1950 census. Also, many population tabulations are available separately for the residence groups of population—urban, rural-nonfarm, and rural-farm—within counties or states, but in general political units have been the primary basis of classification. It is obvious that measures of demographic characteristics which are averages of heterogeneous sub-units may muddle the description of the distribution of characteristics and, especially of importance here, may conceal the existence of relationships between two or more characteristics.

Let us imagine, for the moment, that the distribution of the crude death rate was known for every infinitesimal segment of the United States and that on the data of the death rate a map was made with lines connecting points with equal death rates analogously to the way lines called isobars connect points with equal barometric pressure on a weather map. Now imagine a grid work of 3,000 squares superimposed over the map to give an idea of the "lump" in which we have to obtain our information on such characteristics by counties. These squares would contain segments varying greatly in death rate. Suppose the averages of the segments in a square (county) were computed for all 3,000 squares (counties) and the measures tabulated as a frequency table. Then suppose the grid were shifted slightly changing the boundaries of the squares (counties) and a new distribution made. An infinite number of positions of the grid are possible; each would probably give a slightly different distribution of death rates in the United States. When different shapes and sizes of the 3,000 divisions (rather than equal squares) are permitted (as in the case of actual counties), there are again an infinitely great number of possibilities. Now one of the sets of patterns of divisions is what we actually have in our 3,000 counties in the United States, and it is these county measures with which we must deal. The possibilities of what sort of distributions we should obtain with some other 3,000 divisions are not only interesting to speculate about but deserve serious study. It seems reasonable to suppose that if the 3,000 divisions were made in such a way as to secure the greatest possible degree of internal homogeneity in demographic characteristics, the correlation coefficients between different demographic characteristics would be higher than those we ordinarily find now. The point of other possible divisions is made here, however, only to illustrate the arbitrary nature of the size, shape, and location of the "lumps" in which we obtain our demographic data.

---

[3] See "State Economic Areas of the United States,"Series Census-BAE, No. 15, August 3, 1950; and Donald J. Bogue, *State Economic Areas: A description of the procedure used in making a functional grouping of the counties of the United States*, (Washington: Government Printing Office, 1951).

Size of units. The plotting of the crude death rate by infinitesimally small segments referred to above is of course impossible, because ratios, rates, and other per capita demographic measures have meaning only for a *population*, that is, for a large enough number of people to provide a base to which events or composite frequencies may be related. There is a minimum limit to a population unit below which the conventional population measures have little meaning since they fluctuate erratically if the base is too small. The matter of choice of the *order* of demographic unit—states, counties, townships—should be governed by the nature of the characteristic[4] and the nature of the inquiry, although often it is determined in advance by the order of unit for which data are available. The student of demographic problems is often in the dilemma of choosing between counties which may be too small for stable rates and states which, because they are so large, include very diverse areas. The use of subregions and economic areas will undoubtedly be tried much more extensively since the census has adopted economic areas for the tabulation of some of the 1950 data. The two most important criteria to be considered in choice of the order of unit are homogeneity and sufficient size for stability of rates.

Unequal size of units. Even if in a sociologist's Utopia homogeneous areas large enough for stable rates were delineated by super social scientists and demographic data were collected and tabulated for these areas, demographic statisticians would still not be satisfied with measures of characteristics on these areas for purposes of correlation and other types of analysis in the statistics of relationship. For the areas, if delineated by the criterion of internal homogeneity, would not be of equal size nor contain equal numbers of people, while correlation analysis has been designed for studying relationship between characteristics as distributed among units of equal importance. There have been developed rather elaborate methods of weighting, but these become too involved to be applied easily in multiple and partial correlation analysis, which seek to investigate the aspects of association among a number of factors simultaneously. The inequality of units is usually ignored in practice, although sometimes subunits are recombined into approximately equal units before proceeding with correlation analysis.[5] Usually correlations are computed for characteristics distributed among the 48 states with the observation on the degree of incidence of the characteristic in New York State with its 1950 population of 14,830,192 given just as much weight as the observation for Nevada with its population of 160,083.

---

[4] See William Edwards Deming, *Some Theory of Sampling* (New York: Wiley, 1950), pp. 189–212.

[5] Frederick E. Croxton and Dudley J. Cowden, *Applied General Statistics* (New York: Prentice-Hall, 1939), pp. 740–741.

Advisability of use of correlation in analyzing demographic data. Let us summarize some of the shortcomings of the research situation confronting a population student who wishes to use correlation or other types of analysis in statistics of relationship. The data are offered to him in arbitrarily delimited lumps, of unequal size, for an ill-defined order of unit (for instance, what is the order of the entity which possesses a "death rate"?) and on characteristics which are not distributed normally. In all of these respects the research situation fails to meet the criteria for applicability of many of the methods of analysis. What should be done then? Should one ignore the failings of the data and proceed as if they were not there? Or should one refuse to use the well-developed tools of correlation analysis or analysis of covariance on such imperfect data? Various demographic statisticians answer the question in various ways, usually somewhere between these two extreme practices. The only advice the writer can offer is rather general. One should recognize the imperfections of the data, face them, strive to overcome them wherever possible, or attempt to construct tools that will be more appropriate; but in the absence of perfectly fitting tools and situations one should use whatever tools come nearest to fitting the situation for all the information they can divulge with a constant checking of results obtained by comparison with results obtained by other methods.

Contingency

Definition. Contingency, according to dictionary definitions, can have either of two opposite meanings: (1) the state of being accidental or (2) the state of being dependent upon another event or situation. In statistics we adopt the latter meaning with a slight modification. The *Statistical Dictionary* defines "contingent" as "related in other than a chance manner; associated in other than a random manner." [1] Thus, statistical usage does not ascribe independence or precedence to either one of the characteristics whose association is being investigated by the methods grouped under the title "contingency."

Methods included under "contingency." Although in its broader sense contingency may be used to refer to the state of dependence or association between any two characteristics, the term has often been used in a narrower sense to designate the methods for investigating association between nonquantitative characteristics. In order to differentiate between the several bodies of method for investigating association, we shall adopt this narrower usage of contingency. However, the methods of contingency, developed primarily for the analysis and description of association between nonquantitative characteristics may be used also for investigating association between quantitative characteristics, although they are not the most precise methods available for the latter task. Since the methods we group under "contingency," primarily developed for nonquantitative characteristics, are usually rougher and simpler than the more elaborate methods for quantitative characteristics, we begin the study of the statistics of relationship with this group of methods. And since association between quantitative characteristics can be treated much more precisely and elaborately by other methods, our emphasis in illustrations of application of these methods will be on nonquantitative characteristics.

Utility of methods. The methods of contingency are probably more useful in sociological research than in other more advanced fields of re-

---

[1] Albert K. Kurtz and Harold A. Edgerton, *Statistical Dictionary of Terms and Symbols* (New York: Wiley, 1939), p. 36.

search where devices for precise measurement have been developed to afford data of the quantitative variable type on a greater proportion of the phenomena of interest. Moreover, many of the bases of classification in which the sociologist is interested are not of a quantitative nature and will never be so measured. Geographical classifications utilized in regional research and type of farming areas utilized in rural research are important examples of this sort. Many other classifications, such as sex, race, type of government, marital status, etc., are essentially nonquantitative, and for the investigation of the relationship between any two of these types of characteristics the methods of contingency are appropriate.

Notation. To facilitate the manipulation of distributions of nonquantitative characteristics and to make possible the formulation of procedures, we shall review and expand the system of notation already introduced in Chapters 7 and 15. The simplest sort of nonquantitative classification is a dichotomous one where every unit in a series may be classified either as possessing a certain attribute or as not possessing it. The customary notation already explained involves the assigning of a letter to the attribute, letting this letter written in Roman type enclosed in parentheses represent the number of individuals in a group possessing that characteristic, and letting the same letter written in Greek type enclosed in parentheses represent the number of individuals not possessing the attribute. For example A may be used to designate the attribute of being "white," in which case α is used to designate the attribute of being "not white." Then a group of N individuals classified into two categories, "white" and "not white," may be counted and the results of the enumeration symbolized in this notation as,

$$(A) + (\alpha) = N$$

This notation is useful for any dichotomous, mutually exclusive set of categories. For manifold classifications involving more than two categories subscripts rather than Greek letters are used. The approved notation involves the assigning of a letter to the whole set of categories and the assigning of a subscript number to each category (the number not being indicative of any hierarchical arrangement, however). For example, if a group of N individuals are to be classified according to race color into the five categories white, black, brown, red, and yellow, the qualitative characteristic, race color, may be assigned the letter A and the different colors of races the subscript numbers 1, 2, 3, 4, 5. The result of classifying and enumerating the individuals as to race color may be symbolized thus:

$$(A_1) + (A_2) + (A_3) + (A_4) + (A_5) = N$$

When two or more nonquantitative characteristics are being considered simultaneously, as in the investigation of association between the

two, the notation becomes more complex. If N individuals can be classified into two categories, (A) and (α), on the basis of possessing or not possessing the attribute A, and if they can also be classified into two categories, (B) and (β), on the basis of possessing or not possessing the attribute B, then they can be cross classified into four groups:

(AB) containing all individuals possessing both A and B,
$(A\beta)$ containing all individuals possessing A but not B,
(αB) containing all individuals possessing B but not A,
$(\alpha\beta)$ containing all individuals possessing neither A nor B.

Such a cross classification is possible only when there is information on each individual relating to attribute A and attribute B. Mere group totals for A and B cannot yield the four classes listed above. This is an important item to remember when preparing data for analysis by the method of contingency.

The notation can be extended for three or more attributes.

$$\text{If } (A) + (\alpha) = N$$
$$(B) + (\beta) = N$$
$$(C) + (\gamma) = N$$

then the N individuals may be divided into eight categories by a three way classification. The symbols representing these eight classes are as follows:

$$(ABC), (AB\gamma), (A\beta C), (A\beta\gamma), (\alpha BC), (\alpha B\gamma), (\alpha\beta C), (\alpha\beta\gamma),$$

with the first symbol representing the number of individuals possessing attributes, A, B, and C, the second symbol representing the number of individuals possessing attributes A and B but not C, and so on. Similarly, extensions can be made for the notation of manifold cross classifications.

DESCRIPTION OF TOTAL ASSOCIATION BETWEEN TWO CHARACTERISTICS (DICHOTOMOUS CLASSIFICATIONS) IN A PARTICULAR GROUP

Contingency tables. We shall explain and illustrate the investigation of the existence, direction, degree, and nature of association of two distributions of nonquantitative characteristics first for the simplest case, that is, for two dichotomous classifications. When data are cross classified for two dichotomous classifications, they are usually entered into what is called a contingency table, which has four cells and marginal totals. The general form of a fourfold, or 2 x 2, contingency table is shown in Table 33. The table is slightly different in form from the presentation

tables we have been using, which have totals at the top and left rather than at the bottom. There are two reasons for the rather general practice of altering the convention for this type of table. In the first place, it is the cross classification into categories which is to be emphasized rather than the totals, and, therefore, the totals may be relegated to less important positions. In the second place, it is to be used as a computation table as well as a presentation table, and computation tables usually have their totals placed as in Table 33. The table is also slightly different from

Table 33. FORM OF A FOURFOLD CONTINGENCY
TABLE

	Attribute		
Attribute	β	B	Total
A............	$(A\beta)$	(AB)	(A)
α............	$(\alpha\beta)$	(αB)	(α)
Total..........	(β)	(B)	N

the form used by G. U. Yule, after which it is patterned,[2] in having the positions of the B and β columns exchanged. The exchange has been made to, emphasize the similarity to correlation tables, especially in dealing with type II characteristics, which may be treated by methods of contingency or of correlation.

Illustration of a contingency table. The meaning of this combination of notation and arrangement is best explained by a concrete illustration. A study of "Demographic Characteristics of Women in *Who's Who*" by Clyde V. Kiser and Nathalie L. Schacter [3] makes it possible to divide the ever-married women in *Who's Who* into the two groups of those married only once and those married more than once. It is also possible to classify these women as to whether they have a college degree or not. If we let A denote the attribute of "having a college degree," then α denotes the attribute of "not having a college degree." Similarly, if we let B denote the attribute "married more than once," then β denotes "married only once." The data on these 1,436 ever-married women in *Who's Who* in 1948, when cross-classified by these categories, can be put into a fourfold contingency table similar to the one illustrated above for the general case, as is shown in Table 34.

[2] G. Udney Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), p. 20.
[3] This illustration is adapted from data in Clyde V. Kiser and Nathalie L. Schacter, "Demographic Characteristics of Women in *Who's Who*," *Milbank Memorial Fund Quarterly*, XXVII (October 1949), p. 422, Table 9.

The data are now in the form suitable for investigation of the various aspects of association between education (as measured by having a college degree or not) and number of times married. Let us consider the general features of a problem in association in order to see what we are trying to investigate and what arrangement of data is required for the investigation. The term "association" is applied to the statistical evidence on relationship between *characteristics*, not between or among individual units enumerated or measured. It is the co-occurrence of the two non-

Table 34. DISTRIBUTION OF 1,436 EVER-MARRIED WOMEN IN "WHO'S WHO" IN 1948 BY EDUCATION AND NUMBER OF TIMES MARRIED

Education	Number of times married		
	Only once	More than once	Total
College degree..............	550	61	611
No college degree...........	681	144	825
Total..................	1,231	205	1,436

Source: Adapted from Clyde V. Kiser and Nathalie L. Schacter, "Demographic Characteristics of Women in *Who's Who,*" *Milbank Memorial Fund Quarterly,* XXVII (October 1949), p. 422, Table 9.

quantitative characteristics, "being married more than once" and "having a college degree," in which we are interested; while the individual units enumerated—the 1,436 women—are for the statistical analysis of the problem merely a medium in which the association of the characteristics can be measured. In order to investigate the association between two characteristics of this type, it is always necessary to have recorded the presence or absence of each trait for each individual unit of a group. There are four possible combinations of traits, AB, αB, $A\beta$, $\alpha\beta$, one of which is recorded for each unit. The more units we have, the fuller will be the documentation of the co-occurrence of the two characteristics. The findings, however, will be a description of the association of characteristics, with the description of the group of units, large or small, acting simply as a specification of where, when, and among what sort of units, the described association exists.

Differentiating carefully between the descriptive and the generalizing functions of statistics, we shall first examine the methods for describing the association for *this particular group* of women. We shall consider the various aspects of the association in the order listed above.

Existence of association. Is there any association between the characteristics of having a college degree and number of times married? This

question can be phrased in several ways. Is the proportion of women in this group married more than once the same for women with college degrees as for all of the women? Is the proportion of women having college degrees the same for the women married more than once as for the women married only once, etc.? The simplest way to answer these questions is to compute a pair of percentages for each question. For the first question we find that 10.0 percent of the women with college degrees were married more than once while 14.3 percent of all the women in the group were married more than once. Or, in answer to the second question, we find that of the women married more than once 29.8 percent of them had college degrees while of all the women in the group 42.5 percent had college degrees.

The fact that in each of the pairs of percentages the percentages differ means that the two characteristics, education and number of times married, are not independent of each other. The answer, then, to the existence of an association between education and number of times married is unequivocally "yes" for this group of 1,436 ever-married women in *Who's Who* in 1948.

Direction of association. When there is no undisputed hierarchy in the nonquantitative classification, we cannot specify direction of association by simply saying it is positive or negative, as is possible in the case of quantitative characteristics. Instead it is necessary to specify that a certain category of one classification is associated positively (or negatively) with a certain category of the other classification. In this particular case we can say that education is negatively associated with number of times married, that is, the more education a woman in this group had the less likely she was to have been married more than once.

Degree of association. By comparing 10.0 percent and 14.3 percent we get some idea of the closeness of the association. However, it is not very satisfactory because it could not be used to compare with other differences in percentages computed from other contingency tables. Several coefficients measuring the degree of association have been devised which are pure numbers, free from the effects of the absolute quantities or of the size of the marginal percentages. One such coefficient is called the "coefficient of association" and is represented by Q. It varies from -1 when there is complete negative association to 0 when there is no association to $+1$ when there is complete positive association. In its range of values and in its meaning it is similar to r, the Pearsonian coefficient of correlation, which measures the degree of association between quantitative variables, but Q is *not* identical with r and not equivalent to it. Q is defined and computed thus,

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \tag{1}$$

For our illustration

$$Q = \frac{(61)(681) - (144)(550)}{(61)(681) + (144)(550)} = -.312$$

Now Q is a measure of degree of intensity or closeness of association, the absolute value of which has more meaning after one becomes familiar with the coefficient. Like certain other statistical coefficients it measures something which cannot be expressed precisely in nonstatistical terms. Its limiting values are more easily understood: if every woman in this group without a college degree had been married only once and every woman with a college degree had been married more than once, then the association would be perfect and Q would have a value of 1. If the situation were reversed, then the association would still be perfect but the value of Q would be -1. If exactly the same proportion of women with college degrees as women without college degrees had been married more than once, Q would have had a value of zero. The sign of Q indicates the direction of the relationship between A and B.

Nature of association. By nature of association we mean the amount or proportion of change in the incidence of one attribute which is associated with the change from one to another of the categories of the other attribute. For nonquantitative characteristics the nature of association cannot be so concisely expressed as in the case of quantitative characteristics where the regression equation describes the nature of the association. Instead we use percentages. The percentages we use depend on what particular feature of the association we are interested in. We might say that 10.0 percent of the women with college degrees were married more than once while 17.5 percent of the women without college degrees had been married more than once. In selecting percentages to be compared, one should remember that any cell entry may be expressed as a percentage of either its row marginal total or its column marginal total and that any row or column marginal total may be expressed as a percentage of the grand total.

Summary of the description of association. To summarize, for this group of 1,436 ever-married women in *Who's Who* in 1948, we have investigated and described four aspects of the association observed between education and number of times married. The information may be condensed in the form shown on page 363.

DESCRIPTION OF TOTAL ASSOCIATION BETWEEN TWO CHARACTERISTICS (DICHOTOMOUS CLASSIFICATIONS) IN A UNIVERSE

In many sociological studies nothing more in the analysis of relationship is attempted than such a description of association for the sample as

SUMMARY OF DESCRIPTION OF ASSOCIATION BETWEEN EDUCATION AND
NUMBER OF TIMES MARRIED FOR 1,436 EVER-MARRIED WOMEN IN "WHO'S
WHO" IN 1948

Aspect of association	Description
1. *Existence of association*	The characteristics are associated.
2. *Direction of association*	The association is negative between "having college degree" and "married more than once."
3. *Degree of association*	The association is of "moderate degree" as measured by $Q = -.312$.
4. *Nature of association*	(Several alternatives here.) The percentage of women with college degrees that were married more than once is 10.0 percent while 17.5 percent of the women without college degrees had been married more than once.

is given above. This is often the case when the study is made for the purpose of learning about a group for a better understanding of their particular situation, perhaps with a view toward changes in policy or administration for that group. When a study utilizes only the descriptive function of statistics, its validity is not dependent upon number of cases studied or upon method of sampling. The description is accurate (if the data are accurate) for the particular group studied, and there is no need for measures of error and precision or for tests of significance; in fact, these are out of place in a purely descriptive study which does not seek to generalize. Many of the studies made under the classification of "social surveys" are of this nature.

Existence of association. Although for the 1,436 women the answer to the question of the existence of association between education and number of times married is seen to be unequivocally yes by a mere inspection of two percentages, for the universe of possibilities, from which this group may be considered a random sample, a more careful test is required. The sample shows over half again as great a proportion of women without college degrees married more than once as is true for women with college degrees, but may not this difference in proportions have been observed in a sample of 1,436 women even though the proportions were the same in the infinite universe from which the sample was drawn? A statistical test of significance is necessary to answer the question. The null hypothesis in this case is that there is *no* association between education and number of times married in this infinite hypothetical universe.

There are two kinds of tests we can make of the null hypothesis here. The first is to test, by the methods in Chapter 19, the significance of the

difference between two corresponding percentages or proportions which, under the null hypothesis, we should expect to be equal. The second is to test the significance of the difference between the observed frequency distribution in the four cells of the contingency table and the frequency distribution expected under the null hypothesis. We shall illustrate the second type test since the first type is illustrated in Chapter 19.

The general procedure in making the second test is to compute the distribution by education and number of times married that we would have expected under the null hypothesis, to compare our observed distribution with this expected distribution by computing chi square, and

Table 35. DISTRIBUTION OF 1,436 EVER-MARRIED WOMEN IN "WHO'S WHO," 1948, BY EDUCATION AND NUMBER OF TIMES MARRIED (SEPARATELY)

Education	Number of times married		Total	
	Only once	More than once	Number	Percent
College degree.............	(1)	(3)	611	42.5
No college degree...........	(2)	(4)	825	57.5
Total: Number.............	1,231	205	1,436	100.0
Percent.............	85.7	14.3	100.0	

Source: Table 34.

to see if the discrepancies between the two distributions are greater than could be explained by the chance variation which occurs in random sampling.

We can calculate how many women would fall into each of the cells of the fourfold contingency table under the null hypothesis by simply knowing the marginal totals. These expected cell frequencies we call "independence" values because they are the numbers which would be expected in the case of complete independence between education and number of times married. The calculation of these independence values, or expected frequencies, is straightforward. Looking at Table 35 we see that 14.3 percent of all the women had been married more than once. Therefore, if education and number of times married are independent, we would expect 14.3 percent of the women with degrees to be married more than once and the same percentage of women without degrees to be married more than once. Looking at it another way, since 42.5 percent of all the women had degrees, we would expect 42.5 percent of the married only once women to have degrees and 42.5 percent of the married more than once women to have degrees. These two approaches give the same expected frequencies. In actual practice we do not have to compute the

marginal percentages. To obtain the expected frequency in a particular row and column, we multiply that row total by that column total and divide by the grand total. In a 2 x 2 table we actually have to compute the expected frequency of only one cell. When this is calculated, the expected frequencies of the other cells can be obtained by subtraction from the row and column totals.

Computing the expected frequency for the number of women without degrees married only once, which we symbolize as $(\alpha\beta)_c$, we get

$$(\alpha\beta)_c = \frac{(1,231)(825)}{(1,436)} = 707.2$$

Table 36 shows the expected frequencies for all the cells when education and number of times married are independent.

Table 36. EXPECTED DISTRIBUTION OF 1,436 EVER-MARRIED WOMEN IN "WHO'S WHO," 1948, BY EDUCATION AND NUMBER OF TIMES MARRIED UNDER THE HYPOTHESIS OF INDEPENDENCE

Education	Number of times married		Total
	Only once	More than once	
College degree.........	523.8	87.2	611
No college degree.......	707.2	117.8	825
Total..............	1,231	205	1,436

Source: Tables 34 and 35.

Now we have the data with which to make a statistical test of significance. Here we want to test the significance of the difference between the distribution expected under the hypothesis of independence (Table 36) and the distribution actually observed (Table 34). The test we shall use is the chi square test, the same one used in the analysis of the distribution of the frequencies of the class intervals of quantitative characteristics to compare an observed distribution with an expected distribution. The chi square distribution is the sampling distribution of the sums of independent squares, and we, therefore, compute chi square as in the previous case, by summing the squares of the deviations of the observed from expected frequencies for each cell, divided by the expected frequency for that cell. If we let f represent the observed frequency in each cell and f_c the expected frequency in each cell, then,

$$\chi^2 = \sum \frac{(f - f_c)^2}{f_c} \tag{2}$$

For our illustration the computations for chi square are shown in Table 37, and we get $\chi^2 = 15.98$.

To find the probability that a distribution yielding so large a chi square would have been observed in a sample of 1,436 if the universe from which the sample was drawn has no association between education and number of times married, we must determine the number of degrees of freedom associated with this chi square. Since the expected distribution was computed from the marginal totals, there is only one degree of free-

Table 37. COMPUTATIONS FOR OBTAINING CHI SQUARE FROM THE FREQUENCIES OF TABLES 34 AND 36

Cell number	f	f_c	$(f - f_c)$	$\dfrac{(f - f_c)^2}{f_c}$	$\dfrac{f^2}{f_c}$
(1)	(2)	(3)	(4)	(5)	(6)
(1)	550	523.8	26.2	1.31	577.51
(2)	681	707.2	−26.2	.97	655.77
(3)	61	87.2	−26.2	7.87	42.67
(4)	144	117.8	26.2	5.83	176.03
Sums	1,436	1,436	0.0	15.98	1,451.98 −1,436.00
					15.98

dom. This can be seen from the fact that it was necessary to compute only one expected frequency and the rest could be immediately determined by subtraction from the marginal frequencies. It is not always true that the chi square for a fourfold table has only one degree of freedom associated with it. Sometimes the expected frequencies may be theoretically determined instead of determined from the marginal totals. In such a case only the grand totals have to agree, and there are three degrees instead of one degree of freedom.

In Appendix Table E we find the chi square value for one degree of freedom and for a probability value of .001 is 10.827. This means that deviations from expected frequencies caused by chance variations of sampling would be so large as to produce a chi square of 10.827 or more only one in 1,000 times. Our chi square of 15.98 is greater than 10.827, and, therefore, if the universe has no association between education and number of times married, the probability is less than .001 that a sample would have been observed with frequencies deviating from the expected frequencies as much as those of our sample. This is so small a probability that we deem the null hypothesis untenable and reject it. Interpreted in

terms of the data, this means that the association between education and number of times married observed in our sample of 1,436 women is so marked that we cannot believe the two characteristics are not associated in the universe from which our sample was drawn. We answer the question as to the existence of association between the characteristics in the universe first with a statistician's double negative and finally with a sociologist's "yes"—that is, our results show that there *is* an association between education and number of times married for the infinite universe of possibilities from which our 1,436 cases are considered a random sample.

Direction of Association. If the question as to the existence of association in the universe has been answered affirmatively, then any of the methods suggested for determining and describing the direction of association for the sample may be used to determine and describe the direction of association in the universe.

Degree of Association. The coefficient of association already described, Q, has been found to be $-.312$ for the sample. This value may also be used as an estimate of the coefficient of association for the universe. The chi square test applied to ascertain the existence of association in the universe is at the same time a test of the "significance" of the coefficient of association. That is, the determination of the existence of association in the universe also proves that Q is significantly different from zero.

Nature of association. Any of the several percentages or any of the differences between pairs of percentages computed for the sample may be used as estimates of the corresponding universe percentages or differences between percentages. By the methods set forth in Chapters 15 and 18 we can set up confidence limits of the estimates and can test the significance of differences observed between sample percentages.

For describing the nature of the association between education and number of times married in the hypothetical universe, let us choose the two percentages, 44.7 percent of the women married only once who have college degrees and 29.8 percent of the women married more than once who have college degrees. We estimate that the corresponding universe percentages have the same values and then set up confidence limits for the estimates by the approximate method explained in Chapter 15.

The 95-percent confidence limits for the first percentage are

$$p_{1,2} = .447 \pm 1.96 \sqrt{\frac{(.447)(.553)}{1,231}} = .418 \text{ and } .476$$

The 95-percent confidence limits for the second percentage are

$$p_{1,2} = .298 \pm 1.96 \sqrt{\frac{(.298)(.702)}{205}} = .235 \text{ and } .361$$

The work of computing the above formulas is shortened tremendously if one uses the fourth column of Table 1 in *The Kelley Statistical Tables*[4] and the fifth column of *Barlow's Tables*.[5]

Figure 37.   Representation on Percentage Scale of Estimates of Two Universe Percentages with Their 95-percent Confidence Limits.

These two estimates with their 95-percent confidence limits are shown in Figure 37. The fact that the confidence limits do not overlap means that the difference between the two is significant beyond the .05 level. Of course, the chi square test for existence of association between education and number of times married has already proved the same thing.

SUMMARY OF DESCRIPTION OF ASSOCIATION BETWEEN EDUCATION AND NUMBER OF TIMES MARRIED USING THE 1,436 EVER-MARRIED WOMEN IN "WHO'S WHO" IN 1948 AS A SAMPLE

Aspect of association	Description
1. *Existence of association*	The characteristics are associated—chi square test demonstrates significance of observed association.
2. *Direction of association*	The association is negative between "having college degree" and "married more than once."
3. *Degree of association*	$Q = -.312$. (No measure of the precision of Q was computed.)
4. *Nature of association*	(Several alternatives here.) The percent of married only once women with college degrees is 44.7 percent with 95-percent confidence limits of 41.8 percent and 47.6 percent; the percent of married more than once women with college degrees is 20.8 percent with 95-percent confidence limits of 23.5 percent and 36.1 percent.

[4] Truman Lee Kelley, *The Kelley Statistical Tables*, rev. ed. (New York: Macmillan, 1948).
[5] L. J. Comrie (ed), *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots, and Reciprocals of All Integer Numbers up to 10,000*, 3d. ed. (London: E. and F. N. Spon, 1935).

Another **method of computing chi square.** Chi square has been defined as

$$x^2 = \sum \frac{(f - f_c)^2}{f_c} \tag{2}$$

expanding the numerator of the right member of the equation,

$$x^2 = \sum \left(\frac{f^2 - 2ff_c + f_c^2}{f_c} \right)$$

dividing the fraction into parts and simplifying,

$$x^2 = \sum \left(\frac{f^2}{f_c} - \frac{2ff_c}{f_c} + \frac{f_c^2}{f_c} \right) = \sum \left(\frac{f^2}{f_c} - 2f + f_c \right)$$

summing the parts separately,

$$x^2 = \sum \frac{f^2}{f_c} - 2\Sigma f + \Sigma f_c = \sum \frac{f^2}{f_c} - 2N + N$$

we finally have the formula to be used in computing,

$$x^2 = \sum \frac{f^2}{f_c} - N \tag{3}$$

It is evident that $\sum \frac{f^2}{f_c}$ is more easily computed than $\sum \frac{(f - f_c)^2}{f_c}$, since

no subtractions have to be made. $\sum \frac{f^2}{f_c}$ is usually denoted by the single

letter S, which makes the formula for chi square

$$x^2 = S - N \tag{4}$$

Column (6) of Table 37 shows the computation of chi square for our example by this method.

Along with methods of computing chi square should be presented a method of "correcting for continuity" which must be used when the cell frequencies are small. This method is applicable for any chi square test where there is only one degree of freedom. It should be used, according to

George W. Snedecor,[6] whenever the expected frequency in any cell of the table is less than 200; or according to R. A. Fisher and F. Yates,[7] whenever the expected frequency in any cell of the table is less than 500. The correction is important when the value of chi square is near the significance level since the correction always reduces chi square and failure to apply the correction might lead one to judge as significant a chi square which should not be so considered.

The correction consists in reducing by one-half unit every deviation of an observed from an expected cell frequency. Applied to column (4) of Table 37 it would reduce the absolute value of these entries from 26.2 to 25.7. This would yield a corrected chi square of 15.37, an unimportant reduction in this case since chi square is so far beyond the .001 significance level.

Other measures of degree of association. There are coefficients other than Q which measure degree of association between nonquantitative characteristics. We will comment briefly on several of these, give formulas for their computation, and compute their value for our example.

The coefficient of mean square contingency was developed by Pearson and is given the symbol C. Its upper limit for a fourfold table is .707, and as the number of cells in the table increases, its upper limit increases, being .894 for a 5 x 5 table. Yule recommends not using C for any table smaller than a 5 x 5 one, although this is frequently done.[8]

$$C = \sqrt{\frac{x^2}{x^2 + N}} \tag{5}$$

For our example this gives a value of

$$C = \sqrt{\frac{15.98}{15.98 + 1,436}} = .105$$

Another coefficient which measures the degree of association is T, which has an upper limit of 1.0 no matter how many or how few cells there are in the table. It is defined in terms of an intermediate coefficient, phi square.

$$\phi^2 = \frac{x^2}{N} \tag{6}$$

---

[6] George W. Snedecor, *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), p. 22.

[7] R. A. Fisher and F. Yates, *Statistical Tables for Biological Agricultural and Medical Research* (London: Oliver and Boyd, 1938), p. 3.

[8] Yule shows that the maximum limit of C in a t x t-fold table is $C = \sqrt{\frac{t-1}{t}}$. *Op. cit.*, p. 54.

$$T^2 = \frac{\phi^2}{\sqrt{(s-1)(t-1)}} \tag{7}$$

where s equals the number of rows and t equals the number of columns. From (7) it is obvious that for a 2 x 2 table $\phi^2 = T^2$. In our example

$$T = \sqrt{\frac{15.98}{1,436}} = .106$$

If it can be assumed that A and B are really quantitative variables that have been arbitrarily dichotomized into A and α, and B and β, then it is possible to compute a product moment coefficient of correlation, r. It can be shown that in this case r is equivalent to T. In this situation it is also possible to compute a tetrachoric correlation, r_t, which will be an approximation to r.[9]

DESCRIPTION OF TOTAL ASSOCIATION BETWEEN TWO CHARACTERISTICS (MANIFOLD CLASSIFICATIONS)

Aspects of association investigated dependent on type of characteristic. Some of the methods presented for analyzing and describing association between characteristics with only two categories can be used when there are more than two categories in either or both of the classifications. Whether or not all of the *aspects* of association have meaning in such a problem depends on the types of the two characteristics. If for each characteristic the categories are differentiated on a strictly non-quantitative basis, the aspect of *direction* of association has no meaning for the distributions considered as wholes, although for any pair of categories (one of each characteristic) the direction may be interpreted as in the case of dichotomous classifications. For example, if for the counties of the United States we analyzed the association between the characteristic "regional location" (with six major regions as categories) and the characteristic "most important source of income" (with categories— agriculture, mining, industry, trade, and transportation), we could investigate existence of association and degree of association between the two characteristics but not the direction and only a modified concept of nature of association. If, however, the categories of the characteristics are ordered, as in the case of type II A characteristics, direction and nature of association have their ordinary meaning. This is the type of example we shall choose to illustrate the analysis and description of association between two characteristics with manifold classifications.

[9] Charles C. Peters and Walter R. Van Voorhis, *Statistical Procedures and their Mathematical Bases* (New York: McGraw-Hill, 1940), pp. 366-375.

Arrangement of ordered manifold categories into a contingency table. Table 38 shows the distribution of 2,957 youths by two characteristics which are treated as if they were nonquantitative although actually their "categories" are ordered. The ordering of the categories of the characteristic "education completed by father" is self-explanatory; we may note that in such a case when cross-tabulating we place at the top of the stub the category which can be considered as the "highest" class interval of

Table 38. DISTRIBUTION OF 2,957 WHITE HIGH SCHOOL YOUTHS OF DURHAM, NORTH CAROLINA, BY EDUCATION OF FATHER AND MARITAL STATUS OF PARENTS, 1939

Education completed by father	Marital status of parents *			Total
	Divorced	Separated	Living together	
College................	19	25	573	617
High school............	21	12	443	476
Part of high school......	12	23	725	760
Sixth or seventh grade...	6	12	700	718
Less than sixth grade....	2	0	384	386
Total.................	60	72	2,825	2,957

* Data are given only for youths with both parents living.
Source: Unpublished data from North Carolina Youth Survey, Cooperative Personnel Study, University of North Carolina.

the characteristic. The characteristic "marital status of parents" is not necessarily ordered, but if we regard it as "marital stability of parents," the three categories of the caption can be thought of as three progressive degrees of marital stability. The categories in the caption are arranged with the lowest on the left, progressing to the highest on the right. The arrangement shown here is by no means adhered to inflexibly for presentation, but it is strongly recommended for analysis in order that the tabular presentation of direction of association will correspond with that of a correlation table. The totals might just as well be placed at top and left, but we have them here at bottom and right to show the similarity to the contingency table.

Computation of χ^2 and C. As a first step in the analysis of association between the two characteristics, we shall compute chi square to investigate the *existence* of association. The method of computing chi square involves the computation of an intermediate measure, P, and use of the relationship

$$\chi^2 = NP - N \qquad (8)$$

To avoid a cumbersome notation we shall not express the operations for obtaining P by a formula but shall simply describe them and trust that the student can follow the computations shown in Table 39. The com-

Table 39. COMPUTATIONS FOR OBTAINING CHI SQUARE AND C FROM THE FREQUENCIES IN TABLE 38 USING P AS AN INTERMEDIATE VALUE

Cell number	f (1)	f^2 (2)	$\dfrac{f^2}{\text{row total}}$ (3)	Sum of (3) for each column (4)	$\dfrac{(4)}{\text{column total}}$ (5)
1	19	361	.5851		
2	21	441	.9265		
3	12	144	.1895	1.7616	.0294
4	6	36	.0501		
5	2	4	.0104		
6	25	625	1.0130		
7	12	144	.3025		
8	23	529	.6961	2.2122	0307
9	12	144	.2006		
10	0	0	.0000		
11	573	328,329	532.1378		
12	443	196,249	412.2878		
13	725	525,625	691.6118	2,700.4991	.9559
14	700	490,000	682.4513		
15	384	147,456	382.0104		
					$P = 1.0160$

putations are made by columns, thus: the frequency in each cell is squared and divided by its row total, the resulting quotients for each column are summed and divided by the respective column total, and the final column quotients summed to obtain P.

Substituting in formula (8) the value of P obtained by these operations and 2,957 for N, we have

$$\chi^2 = (1.0160)(2,957) - 2.957 = 47.31$$

Substituting this value into (5) we get

$$C = \sqrt{\frac{47.31}{47.31 + 2,957}} = .13$$

Degrees of freedom. To answer the question of existence of association in the universe we must know the number of degrees of freedom

associated with χ^2. If s is the number of rows, and t is the number of columns, then when the expected distribution has been determined from marginal totals, the number of degrees of freedom, *d.f.*, is given by the formula,

$$d.f. = (s - 1)(t - 1) \tag{9}$$

In our example,

$$d.f. = (5 - 1)(3 - 1) = 8$$

The reasonableness of formula (9) can be seen without any advanced theoretical treatment of degrees of freedom. It must be remembered that while we have used a short-cut method of computation for χ^2 which did not require our computing explicitly the "independence values" of the cell, still we have computed chi square as a measure of departure of observed frequencies from "expected" frequencies determined by marginal totals. It may help to imagine Table 38 as if it were similar to a crossword puzzle, vacant of cell entries with only marginal totals showing. Now consider how much freedom we would have in filling up the cell entries subject to the stipulation that the entries must add up to the marginal totals shown. Beginning with the first column, we could place any frequency (less than its row or marginal total) in the top cell and any frequencies in the next three cells of the first column (so long as their sum is less than the column total). If the frequencies of all five cells of column (1) must add up to 60, we have no freedom of choice left at all when we get to the fifth cell, for it must be the value which added to the sum of the other four frequencies will make 60, the column total. For each cell entry where we can choose any frequency we wish, (as long as it is not too large to run over totals) we say we have one degree of freedom. Thus, in filling the five cells of the first column we have four degrees of freedom. Similarly, for filling the five cells of the second column we have four degrees of freedom. But when we start on the third column, again we have no choice because two cells in the first row are already filled and the frequency of the third cell (first row and third column) must be the value which added to the sum of the other two frequencies will make 617, the row total. The same situation holds for the other entries of the third column. Thus, we see that for all except one column we have as many degrees of freedom less one as there are rows, and this is exactly what formula (9) expresses.

Existence and degree of association. From Appendix Table E we find that the probability of observing a χ^2 as large as 47.31 with eight degrees of freedom is less than .001. Therefore, we reject the hypothesis that there is no association between the two characteristics in the universe. The chi square test tells us that the observed coefficient of con-

tingency, .13, is significantly greater than zero, and we shall use it as an estimate of degree of association in the universe.

Direction of association. When the square root is extracted to obtain C, there is no way of knowing whether to take the positive or the negative root. We shall have to obtain information on direction of association in some other way. There are several choices. The simplest is to express all the frequencies in one row as component percentages of the row total and then to note whether the cells indicating the higher levels

Table 40. COMPARISON OF THE DISTRIBUTION OF TABLE 38 WITH THE DISTRIBUTION EXPECTED UNDER THE HYPOTHESIS THAT THERE IS NO ASSOCIATION BETWEEN THE CHARACTERISTICS

Education completed by father	Marital status of parents			
	Divorced	Separated	Living together	Total
College..............	(13) 19 (+) *	(15) 25 (+)	(589) 573 (−)	617
High school..........	(10) 21 (+)	(12) 12 (0)	(455) 443 (−)	476
Part of high school.....	(15) 12 (−)	(18) 23 (+)	(726) 725 (−)	760
Sixth or seventh grade..	(15) 6 (−)	(17) 12 (−)	(686) 700 (+)	718
Less than sixth grade...	(8) 2 (−)	(9) 0 (−)	(369) 384 (+)	386
Total................	60	72	2,825	2,957

* The sign after an observed frequency denotes the direction of its deviation from expectation.

Source: Table 38.

of education of father have higher or lower percentages of parents in the category of highest marital stability than the lower levels of education. For more careful analysis it is perhaps better to compute the independence values of frequencies and compare them with the observed frequencies. As before, the independence value for a cell is computed by multiplying its row total by its column total and dividing the product by the total number of cases. Table 40 shows both expected and observed frequencies for the example with the direction of deviation from observed frequency indicated. The signs of the deviations form a pattern with a band of plus signs extending from upper left to lower right and with minus signs in the upper right and lower left portions. Interpreting the signs as in the case of dichotomous classification we can say for any one category of marital status what the direction of association is with any category of education.

Since the categories are ordered, however, and since the pattern of signs is consistent, we can designate the direction of association between the two characteristics "education of father" and "marital stability of parents" as *negative* because the deviations of frequencies have a negative sign in high categories of both and in low categories of both.

Nature of association. The percentages suggested would show certain features of the nature of the association, that is, how much greater percentage of college graduates are divorced than of parents who did not complete the sixth grade. There is no very concise method, however, of summarizing the description of the nature of the association by the simple methods of contingency. If such a description is desired, the methods of correlation must be used.

TOTAL AND PARTIAL ASSOCIATION

Definitions. The association between two characteristics among a group of units considered without any regard for the distribution of other characteristics among those units is called total association. The several coefficients of association computed for education and number of times married among the women in *Who's Who* in 1948 were all coefficients of total association between those characteristics for the universe as defined. When there is information on the distribution of some other characteristic among the units of the same universe, the association between the first two characteristics within categories of the third characteristic is called partial association and is described by various summarizing measures, often called coefficients of partial association.

Limitations of the description of total association. The neglect of other relevant factors may lead to erroneous interpretations of coefficients of total association, especially if one imputes causal significance to the association, that is, if the association is interpreted to indicate a "relationship" as defined in Chapter 20. Suppose in a study similar to the above example one discovers a high coefficient of total association between education and number of times married with women with more education being married fewer times. If the investigation is carried no further, one may be inclined to regard this as a constant relationship. Suppose, however, that the relationship between education and number of times married is not really direct, as it appears, but is indirect, explained by the association of each of these characteristics with the general type of occupation of the women. Women with college education who get in *Who's Who* are likely to be in the scientific, educational, legal, medical, or other professions that in general have relatively conservative codes of personal conduct. Being a member of such a group may tend to inhibit a woman from divorce and remarriage. On the other hand, women

without a college education who get in *Who's Who* are more likely to have attained eminence in the fields of entertainment or the arts, in which divorce and remarriage are more accepted. If this were the case, we could not detect it unless categories of occupation were considered in the analysis. For example, within the group of ever-married women listed in *Who's Who* who have attained recognition in the fields of entertainment and the arts we might discover that there was no association between education and number of times married or that the association was positive. This illustrates the fact that in an analysis of relationship neglecting to consider relevant factors may lead to erroneous interpretation of the results. Let us make this point clear—the coefficient of total association between education and number of times married is not "wrong"; it describes the facts of co-occurrence of these two variables in this group. However, it might be misleading when one is trying to use association as suggestive of a relationship with predictive value, in cases such as the above where the two variables might not be associated within the categories of a third characteristic. Another type of case where a coefficient of total association may be misleading is where it is a sort of an average of two or more quite different coefficients of partial association within the categories of a third characteristic.

Investigation of partial association by subdividing the original universe. In actual research one is often primarily interested in the relationship between two particular characteristics and proceeds to investigate the relationship by describing the association between these two within the categories of another characteristic. The original universe is subdivided into as many subuniverses as there are categories of the third characteristic, and the partial association within each subuniverse is described. There may be a fourth relevant characteristic, in which case the units in each subuniverse are cross-classified by the fourth characteristic and thereby subdivided further into as many subuniverses as there are categories of the fourth characteristic. Finally, the partial association between the first two characteristics is described for each of the subuniverses. Theoretically the number of characteristics considered could be increased indefinitely; actually in sociological research the numbers of cases in the subuniverses, and in the cells of the contingency tables for each subuniverse, become too small to give reliable results when many characteristics are considered.

Relativity of terms. The terms "total" and "partial" association are defined in relation to the original universe sometimes called the "universe of discourse," and they are, therefore, relative terms. For instance, if the original universe in the case mentioned above had been all ever-married women, the universe containing only those ever-married women in *Who's Who* would be considered a subuniverse and the descriptive measures of

the association between education and number of times married within this subuniverse would be considered measures of partial association rather than of total association. The principle involved is this—association is more likely to be interpretable as signifying relationship (that is, as having predictive validity) if it is found to exist in a group of units homogeneous with respect to other relevant factors. We usually achieve approximate homogeneity with regard to certain important factors in delineating the original universe to be investigated. It is an arbitrary decision as to what one delimits as the "original universe." The universe should include units which vary in the characteristics one is primarily interested in investigating and also which vary in any characteristics in which one is secondarily interested. Information on association is possible for only those characteristics with respect to which the enumerated units vary (except in the limiting case of perfect association). Therefore, the definition of the original universe sets limits to the scope of the characteristics whose relationships may be investigated. Once the original universe is defined for a problem, any association between two characteristics within that entire universe is called total association, and any association between two characteristics within subuniverses of that universe is called partial association.

An excellent illustration of partial association can be found in an article by Clark Tibbitts and Samuel A. Stouffer on "Testing the Significance of Comparisons in Sociological Data." [10] This article also gives some formulas that are useful in testing the significance of differences.

SUGGESTED READINGS

Dixon, Wilfrid J., and Massey, Jr., Frank J., *Introduction to Statistical Analysis* (New York: McGraw-Hill, 1951), Chap. 13.

Peatman, John Gray, *Descriptive and Sampling Statistics* (New York: Harper, 1947), Chap. 4.

Robinson, W. S., "Ecological Correlations and the Behavior of Individuals," *American Sociological Review*, 15 (June 1950), pp. 351–357.

Snedecor, George W., *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), Chap. 9.

Yule, G. Udny, and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), pp. 50–65.

---

[10] Clark Tibbitts and Samuel A. Stouffer, "Testing the Significance of Comparisons in Sociological Data," *The American Journal of Sociology*, 40 (November 1934), pp. 357–363. For an excellent example of the reasoning preceding, accompanying, and following actual statistical procedures, which makes such procedures meaningful, the student is urged to read this article.

Analysis of Variance

"MODERN" STATISTICAL METHODS

Definitions. Analysis of variance is, as its name implies, the analyzing or breaking up of variance into portions arising from specified sources and the testing of these portions to discover if they are significantly different. It comprises a basic body of methods which can be applied to many sorts of situations and which contribute to a better understanding of many other methods of investigating variation, such as correlation and regression. Analysis of covariance, similarly, is the breaking up of covariance (a term to be defined in the next chapter) into portions arising from specified sources and the testing of these portions to see if they are significantly different. Analysis of variance and analysis of covariance, together with methods for dealing with data from small samples, are often referred to as "modern" statistical methods.

Origin of modern statistical methods. Although there were earlier theoretical contributions, the date of the beginning of modern statistical methods is often considered as 1908, when W. S. Gosset published an article [1] presenting "Student's Distribution" referred to in Chapters 16 and 19. The application and extension of Gosset's work and the important developments since 1920 have taken place primarily under the leadership of the English statistician R. A. Fisher, the inventor of analysis of variance and covariance. Fisher's name is often used synonymously with the adjective "modern" in designating the body of statistical methods including small sample theory, analysis of variance and covariance, and experimental design.

The field of application in which modern statistical methods arose is that of agricultural biology. Therefore, until about 1940 exposition of the methods of analysis of variance and covariance was available only in literature dealing with the methods on an advanced and theoretical level

---

[1] Student, "On the Probable Error of a Mean," *Biometrika*, 24 (1908), pp. 1–25.

of mathematical statistics or in literature dealing with the application of the methods to agricultural biology. It is a considerable jump from length of eggs in a cuckoo's nest or weight of litters of swine to sociological research problems. Consequently, students in sociological statistics either had to puzzle over examples of the application of analysis of variance and covariance in a field very foreign to their own or to forego learning these methods.

Application of methods to fields other than agricultural biology. In 1940 E. F. Lindquist published a book devoted almost wholly to the exposition of methods of analysis of variance and covariance in educational research.[2] Since this field of application is much nearer to that of the sociologist than is the field of agricultural biology, the book is recommended for students in sociological statistics. In the educational research situation, however, a greater degree of experimental control is available than in most of the research situations of the sociologist. For instance, classes of school children can have teachers or methods assigned to them at random in an educational research problem investigating the efficacy of different teaching methods. In population research, however, people already *are* in a certain region, in a certain socio-economic class, or of a certain religion, and their falling into certain categories of cross-classifications can only be *observed* by the research person.

The March 1947 issue of *Biometrics* (Vol. 3, No. 1) contains two articles on analysis of variance that are especially recommended to the student. They are "The Assumptions Underlying the Analysis of Variance" by Churchill Eisenhart and "Some Consequences when the Assumptions for the Analysis of Variance Are not Satisfied" by W. G. Cochran. Aside from these articles there has been no thorough treatment of the utility of the methods of analysis of variance and of covariance in the analysis of data from purely *observational* situations as differentiated from *experimental*, although some examples of the use of the methods in such situations have been published.[3] It is to be hoped that sociological research will soon have a constructive and critical treatment of the applicability and utility of the methods of analysis of variance and covariance as thorough as that of Lindquist's for educational research.

---

[2] E. F. Lindquist, *Statistical Methods in Educational Research* (Boston: Houghton, 1940).
[3] R. K. Mukherjee and F. K. Girling, "Breton Family and Economic Structure," *Rural Sociology*, 15 (March 1950), pp. 49–62; ——, "Economic Structure in Two Breton Villages," *Rural Sociology*, 14 (December 1949), pp. 295–305; Melvin Seeman, "Skin Color Values in Three All-Negro School Classes," *American Sociological Review*, 11 (June 1946), p. 317; Samuel A. Stouffer, "A Technique for Analyzing Sociological Data Classified in Non-Quantitative Groups," *American Journal of Sociology*, 39 (September 1933), pp. 180–193; Milton Friedman, "The Use of Ranks in Analysis of Variance to Avoid the Assumption of Normality," *Journal of the American Statistical Association*, 32 (December 1937), pp. 675–701.

Purpose and limitations of the presentations of analysis of variance and covariance.[4] This text does not attempt to meet the above need, however. Probably more exploratory research will have to be done before the need can be met. To provide an explanation of the computation technique in a setting familiar to sociology students, several suggestive applications are presented in this chapter and in Chapter 24. The mechanics of computation of analysis of variance and covariance are not difficult, and it is hoped that their explanation by means of sociological illustrations will stimulate the exploratory research needed. It must be emphasized, however, that the illustrations given of the methods of analysis of variance and covariance applied to sociological data are to be taken as suggestive, not as authoritative. As explained, the purposes of including them are as follows: (1) to present what promise to be fruitful techniques in a setting familiar to sociology students; (2) to illustrate the differences between the research situations of the experimentalist where the methods were developed and of the observationist where their validity has not been established; and (3) thereby to stimulate exploratory work on the part of sociologists in discovering the potentialities and limitations of these methods in their own field.

The place of analysis of variance in statistics of relationship. Although analysis of variance in a broad sense applies to any process of segregating variation according to its source, we shall in this chapter use the term to refer to the methods for analyzing and describing association between one quantitative and one or more nonquantitative characteristics. The simplest situation is one where we wish to know if levels of incidence of a quantitative characteristic differ significantly for the several categories of a nonquantitative characteristic. The limiting case of this situation, where there are only two categories in the nonquantitative characteristic, reduces the problem to the testing of the significance of the difference between two means, as treated in Chapter 19. Analysis of variance is not limited to this case of two categories but affords a test of the significance of the difference between [5] the means of many categories at once. The situation next in complexity involves the analysis of associa-

---

[4] This paragraph appeared in substantially the same form in the first edition of this book, published in 1941. There have been a small number of sociologists who have explored the application of analysis of variance and covariance to social research problems in the intervening decade, but little progress has been made. The authors regret that this paragraph required little revision.

[5] Statisticians have modified ordinary grammatical usage in letting "between" refer to more than two classes. They use "between" in such cases instead of "among" to differentiate more sharply between the variation manifested by summarizing measures of classes (calling this "between-class" variation) and the variation manifested by units within classes (calling this "within-class" variation).

of mathematical statistics or in literature dealing with the application of the methods to agricultural biology. It is a considerable jump from length of eggs in a cuckoo's nest or weight of litters of swine to sociological research problems. Consequently, students in sociological statistics either had to puzzle over examples of the application of analysis of variance and covariance in a field very foreign to their own or to forego learning these methods.

Application of methods to fields other than agricultural biology. In 1940 E. F. Lindquist published a book devoted almost wholly to the exposition of methods of analysis of variance and covariance in educational research.[2] Since this field of application is much nearer to that of the sociologist than is the field of agricultural biology, the book is recommended for students in sociological statistics. In the educational research situation, however, a greater degree of experimental control is available than in most of the research situations of the sociologist. For instance, classes of school children can have teachers or methods assigned to them at random in an educational research problem investigating the efficacy of different teaching methods. In population research, however, people already *are* in a certain region, in a certain socio-economic class, or of a certain religion, and their falling into certain categories of cross-classifications can only be *observed* by the research person.

The March 1947 issue of *Biometrics* (Vol. 3, No. 1) contains two articles on analysis of variance that are especially recommended to the student. They are "The Assumptions Underlying the Analysis of Variance" by Churchill Eisenhart and "Some Consequences when the Assumptions for the Analysis of Variance Are not Satisfied" by W. G. Cochran. Aside from these articles there has been no thorough treatment of the utility of the methods of analysis of variance and of covariance in the analysis of data from purely *observational* situations as differentiated from *experimental*, although some examples of the use of the methods in such situations have been published.[3] It is to be hoped that sociological research will soon have a constructive and critical treatment of the applicability and utility of the methods of analysis of variance and covariance as thorough as that of Lindquist's for educational research.

---

[2] E. F. Lindquist, *Statistical Methods in Educational Research* (Boston: Houghton, 1940).
[3] R. K. Mukherjee and F. K. Girling, "Breton Family and Economic Structure," *Rural Sociology*, 15 (March 1950), pp. 49–62; ——, "Economic Structure in Two Breton Villages," *Rural Sociology*, 14 (December 1949), pp. 295–305; Melvin Seeman, "Skin Color Values in Three All-Negro School Classes," *American Sociological Review*, 11 (June 1946), p. 317; Samuel A. Stouffer, "A Technique for Analyzing Sociological Data Classified in Non-Quantitative Groups," *American Journal of Sociology*, 39 (September 1933), pp. 180–193; Milton Friedman, "The Use of Ranks in Analysis of Variance to Avoid the Assumption of Normality," *Journal of the American Statistical Association*, 32 (December 1937), pp. 675–701.

Purpose and limitations of the presentations of analysis of variance and covariance.[4] This text does not attempt to meet the above need, however. Probably more exploratory research will have to be done before the need can be met. To provide an explanation of the computation technique in a setting familiar to sociology students, several suggestive applications are presented in this chapter and in Chapter 24. The mechanics of computation of analysis of variance and covariance are not difficult, and it is hoped that their explanation by means of sociological illustrations will stimulate the exploratory research needed. It must be emphasized, however, that the illustrations given of the methods of analysis of variance and covariance applied to sociological data are to be taken as suggestive, not as authoritative. As explained, the purposes of including them are as follows: (1) to present what promise to be fruitful techniques in a setting familiar to sociology students; (2) to illustrate the differences between the research situations of the experimentalist where the methods were developed and of the observationist where their validity has not been established; and (3) thereby to stimulate exploratory work on the part of sociologists in discovering the potentialities and limitations of these methods in their own field.

The place of analysis of variance in statistics of relationship. Although analysis of variance in a broad sense applies to any process of segregating variation according to its source, we shall in this chapter use the term to refer to the methods for analyzing and describing association between one quantitative and one or more nonquantitative characteristics. The simplest situation is one where we wish to know if levels of incidence of a quantitative characteristic differ significantly for the several categories of a nonquantitative characteristic. The limiting case of this situation, where there are only two categories in the nonquantitative characteristic, reduces the problem to the testing of the significance of the difference between two means, as treated in Chapter 19. Analysis of variance is not limited to this case of two categories but affords a test of the significance of the difference between [5] the means of many categories at once. The situation next in complexity involves the analysis of associa-

---

[4] This paragraph appeared in substantially the same form in the first edition of this book, published in 1941. There have been a small number of sociologists who have explored the application of analysis of variance and covariance to social research problems in the intervening decade, but little progress has been made. The authors regret that this paragraph required little revision.

[5] Statisticians have modified ordinary grammatical usage in letting "between" refer to more than two classes. They use "between" in such cases instead of "among" to differentiate more sharply between the variation manifested by summarizing measures of classes (calling this "between-class" variation) and the variation manifested by units within classes (calling this "within-class" variation).

tion between one quantitative and two nonquantitative characteristics. Under certain conditions, well known in the experimental situation but little explored in the observational situation, it is possible to test the significance of the difference between the set of means of the categories of one nonquantitative classification allowing for the differences with respect to the other nonquantitative characteristic. Theoretically there is no limit to the number of nonquantitative characteristics which can be considered simultaneously in the analysis of variance.

Qualifications to the interpretation of the examples which follow. In accordance with the purposes of this chapter explained above we shall proceed to the computation techniques for the analysis of variance. Since the sociological problems are based on nonexperimental data, the applicability of some of the procedures we shall illustrate is questionable. Therefore, we remind the reader again that these illustrations must be considered only suggestive.

ANALYSIS OF VARIANCE WITH ONE CRITERION OF CLASSIFICATION

The problem. We shall introduce analysis of variance by means of a simple problem which is typical of many similar ones involving the investigation of regional differentials. In such problems the quantitative characteristic is a measure for states, counties, cities, or other population units. The nonquantitative characteristic is regional location. Rather than investigating the four aspects of association between the quantitative and nonquantitative characteristics in the order they were considered in the chapter on contingency, we shall go immediately into that aspect for which analysis of variance is primarily adapted—the existence of association in the universe. Later we shall give a summary description of the four aspects of association for the sample and for the universe.

For any two regions differences between regional means may be tested by methods already given in Chapter 19, but analysis of variance offers a method of testing significance of differences between the whole set of regional means at once. It does this by separation of the total variation displayed by the units into two parts—the variation arising from the varying of units around their respective regional means and variation arising from the varying of region means around their "grand mean." (The term "grand mean" denotes a mean of means, which is also the mean of all the individual items.) The "sum of squares" (sum of the squared deviations) is actually split up into these two parts, and when the parts are divided by their respective degrees of freedom, two "mean square variances" are obtained. The ratio of the "between-region" variance to the "within-region" variance is called F. Values of F at the .05, .01, and .001 levels of significance can be found in Appendix Table F.

The data. The data for the illustrative problem are taken from the 1950 Census of Population, Preliminary Reports, Series PC-5, giving characteristics of standard metropolitan areas with a population of 250,000 or more in 1940. We have recorded the percentage of females, 14 years of age and over, that were single (never married) in 1950 for each of 18 of these metropolitan areas. This information is shown in Table 41.

Table 41. PERCENTAGE OF FEMALES 14 YEARS OF AGE AND OVER THAT WERE SINGLE IN 18 STANDARD METROPOLITAN AREAS HAVING OVER 250,000 POPULATION, 1950

Region and metropolitan area	Percentage of females, 14 and over, single in 1950
Northeast	
Baltimore, Md.	21
Charleston, W. Va.	19
Johnstown, Pa.	25
Providence, R. I.	26
Springfield–Holyoke, Mass.	27
Wilkes-Barre–Hazleton, Pa.	26
Southeast	
New Orleans, La.	20
Nashville, Tenn.	19
Miami, Fla.	14
Memphis, Tenn.	15
Louisville, Ky.	18
Atlanta, Ga.	18
Middle States	
Youngstown, Ohio	19
Milwaukee, Wis.	23
Indianapolis, Ind.	19
Detroit, Mich.	18
Cleveland, Ohio	19
Chicago, Ill.	20

Source: 1950 Census of Population. Preliminary Reports. Series PC-5.

Sources of variation of the cities in percentage of females married. The term variation is used somewhat generally to refer to the differing of the measures or units in a distribution. Often, however, "variation" is considered to be measured by the "sum of squares"—that is, the sum of the squares of the deviations of the measures of units from their mean. We shall use the term "variation" in this sense. The total variation of these 18 population units in the percentage of females single is the sum of the

squares of the deviations of their measures from the grand mean. By analysis of variance we shall break up this total variation into two parts, one part ascribable to regional differences (between region variation) and the other part ascribable to differences between cities within regions (within region variation).

Table 42. SEPARATION INTO COMPONENT PARTS OF DATA SHOWN IN TABLE 41

	Percentage of females single		
	Northeast	Southeast	Middle States
	21 = 20.3 + 3.7 − 3	20 = 20.3 − 3.0 + 2.7	19 = 20.3 − 0.7 − 0.7
	19 = 20.3 + 3.7 − 5	19 = 20.3 − 3.0 + 1.7	23 = 20.3 − 0.7 + 3.3
	25 = 20.3 + 3.7 + 1	14 = 20.3 − 3.0 − 3.3	19 = 20.3 − 0.7 − 0.7
	26 = 20.3 + 3.7 + 2	15 = 20.3 − 3.0 − 2.3	18 = 20.3 − 0.7 − 1.7
	27 = 20.3 + 3.7 + 3	18 = 20.3 − 3.0 + 0.7	19 = 20.3 − 0.7 − 0.7
	26 = 20.3 + 3.7 + 2	18 = 20.3 − 3.0 + 0.7	20 = 20.3 − 0.7 + 0.3
ΣX	144	104	118
Regional means	24	17.3	19.7
Grand mean	20.3	20.3	20.3
Variations of regional means from grand mean	+3.7	−3.0	−0.7

We can start with the idea that there is some average percentage of females single in each city, and the variations from this average are due to regional effects and individual city variations. Let X be the percentage of females single in a particular city, C_{ij}, in a particular region, R_i. (The subscript i denotes the region, the subscript j denotes a particular city in the region.) We can say then that

$$X = \bar{X} + r_i + c_{ij} \tag{1}$$

where $\bar{X}$ is the average percentage of females single,

$\quad\quad r_i$ is the deviation from average of the ith region, or the region effect, and

$\quad\quad c_{ij}$ is the deviation of the city from the region mean.

Table 42 shows the percentage of females single in each city divided into these three parts. For example, the first value in the Northeast region, 21, is equal to the grand mean, 20.3 plus the deviation of the regional mean from the grand mean, 3.7, plus the deviation of the city from the regional

mean, -3.0. (In some cases rounding prevents the totals from being exact.) Note that the sum of the r_i's is equal to zero, and that the sum of the c_{ij}'s for each column is zero (except for rounding errors).

From equation (1) it is seen that

$$x = r_i + c_{ij} \tag{2}$$

Since the total variation is equal Σx^2, the total variation is also equal to $\Sigma(r_i + c_{ij})^2$. Squaring both sides of equation (2) and summing gives

$$\Sigma x^2 = \Sigma r_i^2 + 2\Sigma r_i c_{ij} + \Sigma c_{ij}^2 \tag{3}$$

Since the summation of all the r_i's is zero and the summation of the c_{ij}'s in each column is zero, it can be shown that the middle term of the expression on the right in equation (3) is identically equal to zero. Therefore,

$$\text{Total variation} = \Sigma x^2 = \Sigma r_i^2 + \Sigma c_{ij}^2 \tag{4}$$

The first of the two terms on the right is the between-region variation, and the second term is the within-region variation. We have thus succeeded in analyzing our variation into these two parts. With the exception of errors due to rounding the student can demonstrate from Table 42 that equation (4) is correct.

In equation (4) we have divided the total variation into two parts, that ascribable to regional differences and the variation due to all other sources. This variation due to all other sources except regions we call our "error" term. This is because we have ascribed it to no specific source. We shall learn later how to subdivide this, ascribe part of it to a specific source, and thus reduce our "error" term.

The task we set ourselves was to determine if the regional differences are significant. We make this test by determining if the variation we have ascribed to regions is significantly different from the error term after we take into consideration the number of degrees of freedom on which each is based. Our null hypothesis is that there is no difference between the variation from the two sources. We test this hypothesis by taking the quotient of the mean square variance estimated from regions and the mean square variance estimated from the error term. This quotient is called F. If the two mean square variances are not significantly different, F is approximately equal to one. The greater F departs from one, the smaller is the probability that we would get such a difference by chance if our data are a sample from a universe in which the two are not different. If this probability is low enough, we conclude that the two mean square variances are significantly different. If the variation due to regions is significantly greater than the variation within regions, we conclude that the regional differences are significant.

If the probability is such that we cannot conclude that the regional

variations are significantly different from the random variations, then either the regional differences are not significant or we have left other major sources of variation in our error term.

Computing the between-class variation and the within-class variation. As is suggested following formula (4), it is possible to compute the between-region (or between-class) variation and the within-region (or within-class) variation from Table 42. If the deviations of regional means from grand means are carried to three decimal places, we compute the variation due to regions as

$$\Sigma r_i{}^2 = 6[(3.667)^2 + (-3.000)^2 + (-0.667)^2] = 137.35$$

The within region variation can be obtained as follows.

$$\Sigma c_{ij}{}^2 = (-3)^2 + (-5)^2 + \cdots + (-0.7)^2 + (0.3)^2 = 94.68$$

This, however, is a very cumbersome way to compute the variation, or sums of squares, as it is frequently called. Let us illustrate a shorter method. First, we compute the total variation, or total sums of squares, Σx^2. For this we need information not shown in Table 42. We need

$$\Sigma X^2 = (21)^2 + (19)^2 + \cdots + (19)^2 + (20)^2 = 7,674$$

We get the total sum of squares from the following formula.

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \tag{5}$$

The second term on the right side of equation (5) is called the correction term, C. We give it a symbol of its own because we use it later. Since $\Sigma X = 366$, we obtain, on substituting in (5),

$$\Sigma x^2 = 7,674 - \frac{(366)^2}{18} = 7,674 - 7,442 = 232$$

This agrees very closely with the sum of the between-class and within-class variation computed directly from Table 42 above.

To compute the between-class variation we use the sums for each region and the number of observations in each region, thus,

$$\text{Between-class variation} = \frac{(144)^2 + (104)^2 + (118)^2}{6} - C = 7,579.33 - 7,442$$
$$= 137.33$$

Now that we have the total variation and the between-class variation computed we calculate the within-class variation by subtracting the between-class variation from the total variation, thus,

$$\frac{\text{Within-class}}{\text{variation}} = 232 - 137.33 = 94.67$$

These computations are summarized in Table 43.

Table 43. COMPUTATIONS FOR OBTAINING SUMS OF SQUARES
FROM THE DATA OF TABLE 41

	Northeast	Southeast	Middle States
	21	20	19
	19	19	23
	25	14	19
	26	15	18
	27	18	19
	26	18	20
Sums	144	104	118

$\Sigma X = 366$ $\Sigma X^2 = 7{,}674$

$$\text{Total variation} = 7{,}674 - \frac{(366)^2}{18} = 7674 - 7442 = 232$$

$$\text{Between-class variation} = \frac{(144)^2 + (104)^2 + (118)^2}{6} - 7{,}442 = 137.33$$

$$\text{Within-class variation} = 232 - 137.33 = 94.67$$

Degrees of freedom. An estimate of variance is always computed by dividing a sum of squares by its associated degrees of freedom. Therefore, the next step is to determine the number of degrees of freedom associated with each of the sums of squares identified in Table 43. In Table 43 we have 18 original observations, but our total variation is the sum of squares of deviations from the mean. The sum of these deviations from the mean must equal zero (see Chapter 8) and imposing this condition uses up one degree of freedom. Thus, the total sum of squares has $(N - 1)$ degrees of freedom. In our example this is $(18 - 1)$ or 17 degrees of freedom.

We saw in Table 42 that the between-class sum of squares can be computed from the deviations of each class mean from the grand mean. We have one of these deviations for each class but again the sum of these deviations must be zero, and this requirement costs us one degree of freedom. Thus, for the between-class variation the number of degrees of freedom is one less than the number of classes. If we have m classes, the number of degrees of freedom for the between-class variation is $(m - 1)$, or in our case, 2.

The degrees of freedom associated with the within-class variation can be obtained by subtraction, that is, by subtracting the degrees of freedom

COMPUTATION GUIDE FOR ANALYSIS OF VARIANCE WITH ONE CRITERION OF CLASSIFICATION AND EQUAL FREQUENCIES

Column

Row	1	2	3	...	j	...	m	Sums
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	X_{1m}	$X_{1\cdot}$
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	X_{2m}	$X_{2\cdot}$
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	X_{3m}	$X_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	X_{i1}	X_{i2}	X_{i3}	...	X_{ij}	...	X_{im}	$X_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
N	X_{n1}	X_{n2}	X_{n3}	...	X_{nj}	...	X_{nm}	$X_{n\cdot}$
Sums	$X_{\cdot1}$	$X_{\cdot2}$	$X_{\cdot3}$	...	$X_{\cdot j}$	...	$X_{\cdot m}$	$X_{\cdot\cdot}$

m = number of classes
n = number of items in each class
$N = nm =$ total number of items

I. Total sum of squares $= \Sigma X_{ij}^2 - \dfrac{(X_{\cdot\cdot})^2}{N}$

degrees of freedom $= N - 1$

II. Between-class sum of squares $= \dfrac{(X_{\cdot1})^2 + (X_{\cdot2})^2 + (X_{\cdot3})^2 + \cdots + (X_{\cdot m})^2}{n} - \dfrac{(X_{\cdot\cdot})^2}{N}$

degrees of freedom $= m - 1$

III. Within-class sum of squares $=$ Total sum of squares minus between-class sum of squares

degrees of freedom $= N - m$

Presentation Table

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total.....................	I	$N - 1$		
Between-class..............	II	$m - 1$	$A = \dfrac{\text{II}}{m - 1}$	$\dfrac{A}{B}$
Within-class.................	III	$N - m$	$B = \dfrac{\text{III}}{N - m}$	

Probability statement

for the between-class variation from the degrees of freedom for the total variation. Symbolized this is $(N - 1) - (m - 1)$ or $(N - m)$. If we consult Table 42, we find that this is a logical result. We computed the within-class variation directly from that table by getting the deviation of each observation from its region mean and summing the squares of these 18 deviations. However, the sum of these deviations must be zero for each region, and we, therefore, lose one degree of freedom for each region, leaving $N - m$, or 15, degrees of freedom associated with the within-class variation.

Mean square variances. The mean square variances can now be computed by dividing each sum of squares by its associated degrees of freedom. The results of these divisions are shown in the fourth column of Table 44. These mean square variances are not additive because they

Table 44. ANALYSIS OF VARIANCE IN PERCENT OF FEMALES SINGLE IN 18 STANDARD METROPOLITAN AREAS BY REGIONAL LOCATION

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	Ratio of variances, F
Total..................	232	17		
Between-region	137.33	2	68.67	
				10.88
Within-region	94.67	15	6.31	

$P[F_{2.15} = 10.88] < .01$

Note: The line above is a short way of saying that the probability of getting an F of 10.88 based on 2 and 15 degrees of freedom is less than .01.
Source: Table 43.

have been computed from different numbers of degrees of freedom. That is why we performed the separation of the variation into parts while it was in the sum of squares stage.

Ratio of variances. The two mean square variances we are interested in comparing are the between-region and the within-region mean square variances. Instead of subtracting one from the other and testing the significance of the difference between them, we use a test of significance based on the ratio between the two. We form the ratio of the between-region to the within-region mean square variance; this ratio is called F.[6] Table 44 shows that the F for our illustration is 10.88. We

[6] The observed value of F, the quotient of the mean square variance estimated from regions divided by the mean square variance estimated from the error term, is generally greater than one. Therefore, Appendix Table F is based only on the upper part of the theoretical distribution of F. The test can still be applied approximately even when the observed F is less than one by comparing the reciprocal of the observed F with the values given in Ap-

compare this value with the values of F in Appendix Table F for .05, .01, and .001 levels of significance. We have degrees of freedom from two sources associated with each F and have to use these degrees of freedom in looking up values in Appendix Table F. The n_1 is the number of degrees of freedom associated with the numerator of the F, and n_2 is the degrees of freedom associated with the denominator of F. In our illustration $n_1 = 2$ and $n_2 = 15$. Frequently we write F_{n_1, n_2} in order to indicate the degrees of freedom associated with the F, as in Table 44 we have written $F_{2,15}$.

In Appendix Table F we find that the F value corresponding to a probability of .001 and an n_1 of 2 and an n_2 of 15 is 11.34. The $F_{2,15}$ corresponding to a probability of .01 is 6.36. Since our F value lies between these, the probability of obtaining the F which we obtained is between .01 and .001.

Meaning of the F test. Let us examine more carefully what is the exact meaning of this test of significance. The two mean square variances are two estimates of variance. The within-region mean square variance is an estimate of the variance in a universe whose units vary about its mean with the dispersion that the 18 population units exhibit around their respective regional means. The between-region mean square variance is an estimate of the variance in a universe whose units vary about its mean with the dispersion which would cause the means of random samples of three units to show the amount of dispersion about their grand mean that the three regional means exhibit about their grand mean. Our test will be to discover if these two estimates of variance could be expected from the same universe. If so, we may say that the variation of region means in this measure is significantly greater than the variation of units within regions, or that the percentage of females single is significantly associated with the regional classification used here, or that regional differentials in percentage of females single are significant.

The hypothesis to be tested is that the two estimates of variance, 68.67 and 6.31, might have been made about the same universe, taking into account the number of degrees of freedom involved in making each

---

pendix Table F. In this case, n_1 and n_2 correspond to the degrees of freedom associated with the larger and the smaller variance respectively. However, if F is statistically significant in this situation interpretation of the results of the test may be difficult.

The F distribution can also be used to test the variances from two different samples for a significant difference. Since neither of the two variances can be regarded as that estimated from the means nor as that estimated from the error term, compute F by dividing the larger variance by the smaller variance and double the value of the probability associated with the tabular value of F with which the computed value is compared. This doubling is necessary because F could have been computed with either of the variances as the numerator. If computed one way, the F is greater than one; if computed the other way, the F is less than one. Therefore, both the upper and lower parts of the F distribution must be taken into account.

estimate. If there were no such things as sampling variation, we should expect F to have a value of one when the hypothesis is true, that is, we should expect our two estimates to be identical and hence the ratio of one to the other would have a value of one. We know, however, that estimates based on only two and 15 independent observations are likely to vary considerably from the universe parameters. The two estimates might both be larger than the parameter, both smaller, or one larger and one smaller. It is the ratio of one of the estimates to the other for which the sampling distribution is described by means of the F table. For the illustration here, the F table tells us the probability is .05 that the estimate based on two degrees of freedom would be 3.68 times as great as (or greater than) the estimate based on 15 degrees of freedom. Similarly, the probability is .01 that the estimate based on two degrees of freedom is 6.36 times as great as our estimate based on 15 degrees of freedom. Since our estimate based on two degrees of freedom is 10.88 times as great as our estimate based on 15 degrees of freedom, we determine that the probability that such an unusual ratio, F, would be observed under the null hypothesis is less than .01. Therefore, we reject the hypothesis and conclude that the variation in percentage of females single arising from the differences between regions is significantly greater than that arising from differences within regions. An analysis of variance is usually summarized and presented as shown in Table 44.

DESCRIPTION OF ASSOCIATION BETWEEN PERCENTAGE OF FEMALES SINGLE AND REGIONAL LOCATION

Existence of association. Any difference in the percentage of females single between regional means indicates that for the sample there is association between this regional classification and the percentage of females single in these 18 standard metropolitan areas. The F test of significance, made through the methods of analysis of variance, indicates that there is association between this regional classification and the percentage of females single in all standard metropolitan areas of these types in these regions (the universe from which this sample is assumed to be randomly drawn).

Direction of association. The direction of association can have no meaning when one of the characteristics is classified into a manifold classification which has no sequential order of classes, as is true in our case. The arrangement or order of the three regions may be changed at will. Therefore, we cannot investigate the direction of association for either the sample or the universe. (In other cases where the manifold classification has a definite hierarchical order, the direction of association does have meaning. The direction can be determined by inspection of

the direction of change in the class means from the first to the last, if such changes are consistent.)

Degree of association. For any given pair of n_1's and n_2's the size of the F of one association may be compared with the size of the F of another association to measure *relative* degree of association. For instance, we can say that the association under investigation here shows a greater degree of association than one which has an F of 5.32, with the same number of degrees of freedom. For comparing two associations where the number of degrees of freedom are different, analysis of variance does not provide comparable measures of degree of association. The comparison of values of P corresponding to the F's affords an inverse comparison of the reliability of our verdict that there *is* association but does not give information on the relative degree of association.

There is a coefficient for measuring the absolute degree of association, which is not so much used now as formerly, called the intraclass coefficient of correlation. The intraclass coefficient of correlation measures the correlation between two series of values, formed by using as corresponding values in the two series each item in a class paired with every other item in that class. If there are n items in a class, there will be $n - 1$ pairs of values for each item, or $n(n - 1)$ pairs for the whole class. Since the same values appear in both series of values, the variances of the two series are identical. This makes for a modification of the ordinary (called "interclass") coefficient of correlation formula. It also sets as a lower limit to the intraclass coefficient $-\dfrac{1}{n-1}$, although the upper limit is plus one. Snedecor gives a formula for the intraclass coefficient of correlation (if all classes have the same number of items) in terms of the mean square variances already computed,[7]

$$\text{Intraclass } r = \frac{V_b - V_w}{V_b + (n - 1)V_w} \tag{6}$$

where V_b = "between-class" mean square variance
V_w = "within-class" mean square variance
n = number of items in each class

For our illustration, substituting the value of V_b, 68.67, and of V_w, 6.31, gives

$$\text{Intraclass } r = \frac{68.67 - 6.31}{68.67 + 5(6.31)} = \frac{62.36}{100.22} = .62$$

This intraclass coefficient of correlation, .62, may be used to describe the

[7] George W. Snedecor, *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1948), p. 243. Also see Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, 1949), p. 230.

degree of association for the sample and also as an estimate of the value of the intraclass coefficient for the universe. That it is significantly different from zero has been tested by the F test.

Still another coefficient for measuring degree of association in analysis of variance is called by Charles C. Peters and Walter R. Van Voorhis the "unbiased correlation ratio," [8] and designated by the symbol ϵ. It is defined by the relation,

$$\epsilon^2 = 1 - \frac{V_w}{V_t} \qquad (7)$$

where V_w = "within-class" mean square variance
and V_t = "total" mean square variance.

Substitution of the values for our illustration, $V_w = 6.31$, and $V_t = 13.65$, gives

$$\epsilon^2 = 1 - \frac{6.31}{13.65} = 1 - .4623 = .5377$$

$$\epsilon = \sqrt{.5377} = .73$$

This ϵ of .73 is a measure of degree of association between percentage of females single and regional location. Its standard error is known, but since the form of its sampling distribution is not known, we cannot test any hypothesis except the hypothesis that the universe value of ϵ is zero. This hypothesis can be tested by tables supplied by Peters and Van Voorhis which give the same result as the F test.[9]

Nature of the association. If the classes of the qualitative classification are not ordered, which is the case of our regional classification, there is no concise way of describing the nature of the association. Perhaps the most commonly used method for describing the nature of the association for the sample is to rank the regions in order of their regional means in the characteristic studied and to present them in this order with their regional means.

Region	Regional mean of percentage of females, 14 years of age and over, single, in 18 standard metropolitan areas
Northeast...	24.00
Middle States...	19.67
Southeast...	17.33

---

[8] Charles C. Peters and Walter R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases* (New York: McGraw-Hill, 1940), pp. 323, 325, 337 ff. Also see Truman Lee Kelley, *Fundamentals of Statistics* (Cambridge: Harvard University Press, 1947), pp. 448–453.
[9] Peters and Van Voorhis, *op. cit.*, pp. 324–325, pp. 494–497, Table 67.

If we wish to describe the nature of the association for the universe—that is, for all standard metropolitan areas with more than 250,000 population in these regions—we use the above regional means as estimates of the universe regional means, but, in addition, we compute confidence limits or other measures of precision. Computation of confidence limits of estimates always involves first calculating the standard error of the statistics which we are using as estimates.

The estimates whose confidence limits we want are means in this case; therefore, the procedure for each region is to compute first an estimate of the standard deviation of the units of that region in percent of females single, since the standard error of the mean is a function of the standard deviation of the distribution. After computing the standard deviation and standard error of the mean for each region, we then set up 99-percent confidence limits about each regional mean. Since the number of cases in each region is so small, only six, we cannot use for obtaining the confidence limits the multiple 2.58 from the table of the normal distribution, but we must take the corresponding multiple from the table of Student's distribution for five degrees of freedom. From Appendix Table D we find this value is 4.032, much larger than the value for the normal distribution. Table 45 shows the computation of the 99-percent confidence limits.

Table 45. COMPUTATIONS FOR OBTAINING 99 PERCENT CONFIDENCE LIMITS OF REGIONAL MEANS OF PERCENT OF FEMALES SINGLE, 1950

Region	ΣX	ΣX^2	$\dfrac{\Sigma X}{N}$ $\overline{X}$	$\dfrac{1}{N}\sqrt{N\Sigma X^2 - (\Sigma X)^2}$ s	$\dfrac{s}{\sqrt{N-1}}$ $\hat{\sigma}_{\bar{x}}$	$\overline{X} \pm 4.032\hat{\sigma}_{\bar{x}}$ 99 percent confidence limits
Northeast....	144	3,508	24.00	2.944	1.317	18.7 and 29.3
Middle States.	118	2,336	19.67	1.599	.715	16.8 and 22.6
Southeast.....	104	1,830	17.33	2.134	.954	13.5 and 21.2

Source: Table 44.

From these 99-percent confidence limits we see that even though the F test has shown that the regional classification divides the population units into classes for which the class means differ significantly as a set, pairs of class means may not differ significantly from each other.

ANALYSIS OF VARIANCE WITH TWO CRITERIA OF CLASSIFICATION

We have analyzed the variance in the percentage of females, 14 years of age and over, single, in 18 metropolitan areas according to one criterion of classification, regions. We will next analyze the variance in the percentage of females single according to another criterion, sex ratio of

SUMMARY OF DESCRIPTION OF ASSOCIATION BETWEEN PERCENTAGE OF
FEMALES, 14 YEARS OF AGE AND OVER, SINGLE, AND REGIONAL LOCATION

Aspect of association	Description for the sample of 18 standard metropolitan areas over 250,000 population	Description for the universe of all standard metropolitan areas over 250,000 population in these regions
1. *Existence of association*	The characteristics are associated—regional means differ.	The association is highly significant—$P[F_{2,15} = 10.88] < .01$.
2. *Direction of association*	This aspect has no meaning in this situation.	This aspect has no meaning in this situation.
3. *Degree of association*	The degree of association is fairly high as measured by Intraclass $r = .62$ or $\epsilon = .73$	The estimates of coefficients measuring degree of association are intraclass $r = .62$, $\epsilon = .73$. $F_{2,15} = 10.88$ may be used to compare the degree of this association with the degree of any other association having an $F_{2,15}$.
4. *Nature of association*		Means in preceding column are estimates of universe means with the following 99-percent confidence limits
Region	Regional mean	
Northeast	24.00	18.7 to 29.3
Middle States	19.67	16.8 to 22.6
Southeast	17.33	13.5 to 21.2

metropolitan areas, then we will make the analysis according to both criteria simultaneously.

Analysis by sex ratio of metropolitan areas. Table 46 shows our 18 standard metropolitan areas divided into two groups depending on whether they have a sex ratio of less than 94 or of 94 and over. Table 47 shows the percentages of females single in each of these cities and the computation of sums of squares following the form for computing shown on page 388. Comparing Tables 47 and 43, we see that the total sums of squares are identical in the two tables, which is to be expected since this is the total variation in the 18 metropolitan areas regardless of how they are classified.

The analysis of variance of these sums of squares is shown in Table 48.

Table 46. CLASSIFICATION OF 18 STANDARD METROPOLITAN AREAS ACCORDING TO WHETHER THE SEX RATIO IS 94 AND OVER OR LESS THAN 94

Group I Sex ratio less than 94	Group II Sex ratio 94 or over
Atlanta, Ga.	Charleston, W. Va.
Baltimore, Md.	Chicago, Ill.
Cleveland, O.	Detroit, Mich.
Indianapolis, Ind.	Johnstown, Pa.
Milwaukee, Wis.	Louisville, Ky.
Nashville, Tenn.	Memphis, Tenn.
New Orleans, La.	Miami, Fla.
Providence, R. I.	Wilkes-Barre–Hazleton, Pa.
Springfield–Holyoke, Mass.	Youngstown, O.

Source: 1950 Census of Population. Preliminary Reports. Series PC-5.

Table 47. COMPUTATION OF SUMS OF SQUARES FROM CLASSIFICATION SHOWN IN TABLE 46 AND DATA OF TABLE 41

	Group I	Group II	
	18	19	
	21	20	
	19	18	
	19	25	
	23	18	
	19	15	
	20	14	
	26	26	
	27	19	
Sums	192	174	$X.. = 366$ $\Sigma X^2 = 7{,}674$

Total variation $= \Sigma X^2 - \dfrac{(X..)^2}{N} = 7{,}674 - \dfrac{(366)^2}{18}$

$\qquad = 7{,}674 - 7442 = 232$

Between-class variation $= \dfrac{(X._1)^2 + (X._2)^2}{n} - \dfrac{(X..)^2}{N} = \dfrac{(192)^2 + (174)^2}{9} - 7{,}442$

$\qquad = 7{,}460 - 7{,}442 = 18$

Within-class variation $=$ Total $-$ between class $= 232 - 18$

$\qquad = 214$

Table 48. ANALYSIS OF VARIANCE IN PERCENTAGE OF FEMALES SINGLE IN 18 STANDARD METROPOLITAN AREAS BY TWO SEX RATIO CLASSES

Source of variation	Sums of squares	Degrees of freedom	Mean square variance	F
Total.....................	232	17		
Between-sex-ratio classes.....	18	1	18	1.35
Within-sex-ratio classes......	214	16	13.375	

$P[F_{1.16} = 1.35] > .05$

Source: Table 47.

Since F is so near one in this case, it is obvious that there is no real difference in variation from these two sources. However, we cannot conclude from this that there is no difference in the percentage of females single between metropolitan areas with low and high sex ratios. It is possible that another major source of variation has not been considered, and when this is true, variation from the unconsidered source, or sources, increases the within-class variation. From our other analysis we know that regional variation is important, and we should take it into consideration. This involves an analysis of variance with two criteria of classification.

Analysis by sex ratio groups and regional groups. When data are cross-classified and we have the same number of cases in each cross-classification, the analysis of variance with two criteria of classification is relatively simple. However, when we have varying numbers of cases in each cross-classification (known as "disproportionate subclass frequencies"), the analysis becomes more complex. We will treat here only the case where the subclasses have equal frequencies.[10] Table 49 shows our 18 metropolitan areas cross-classified by sex ratio groups and regions. From this table we see that there are three cases in each subclass. When we have the same number of cases in each subclass, the sum of squares for one criterion of classification is independent of the sum of squares for the other criterion of classification. Thus, we have divided our total sum of squares into (a), (b), and (c): (a) that part explained by regional differences, (b) that part

---

[10] For methods of handling analysis of variance with disproportionate subclass frequencies see Palmer O. Johnson, *Statistical Methods in Research* (New York: Prentice-Hall, 1949), pp. 260–266; George W. Snedecor and Gertrude M. Cox, *Disproportionate Subclass Numbers in Tables of Multiple Classification* (Iowa State College Experiment Station Research Bulletin 180, 1935); Fei Tsao, "General Solution of the Analysis of Variance and Covariance in the Case of Unequal or Disproportionate Numbers of Observations in the Subclasses," *Psychometrika*, 11 (1946), pp. 107–128; Frank Yates, "The Analysis of Multiple Classifications with Unequal Numbers in the Different Classes," *Journal of the American Statistical Association*, 29 (March 1934), p. 51; R. E. Patterson, "The Use of Adjusting Factors in the Analysis of Data with Disproportionate Subclass Numbers," *Journal of the American Statistical Association*, 41 (September 1946), pp. 334–346. This last article demonstrates a relatively simple approximation technique.

Table 49. CLASSIFICATION OF 18 STANDARD METROPOLITAN AREAS BY REGION AND SEX RATIO, 1950

	Region		
	Northeast	Southeast	Middle States
Sex ratios less than 94......	Baltimore Providence Springfield–Holyoke	New Orleans Nashville Atlanta	Indianapolis Cleveland Milwaukee
Sex ratios 94 and over......	Charleston Johnstown Wilkes-Barre–Hazleton	Miami Memphis Louisville	Youngstown Detroit Chicago

Source: 1950 Census of Population. Preliminary Reports. Series PC-5.

explained by differences between sex ratio classes, and, (c) the part unexplained by either of these sources. We obtain the sum of squares for part (c) by subtracting the sums of squares for (a) and (b) from the total. We call part (c) the discrepance, or error.

Table 50. ANALYSIS OF VARIANCE IN PERCENTAGE OF FEMALES SINGLE IN 18 STANDARD METROPOLITAN AREAS CLASSIFIED BY REGIONS AND SEX RATIO

Source of variation	Sum of squares	Degrees of freedom	Mean square variance
Total.................................	232.00	17	
a. Between region.....................	137.33	2	68.67
b. Between-sex-ratio classes.............	18.00	1	18.00
c. Discrepance.........................	76.67	14	5.48

F's

$$\frac{a}{c} = 12.53; \quad P[F_{2.14} = 12.53] < .001$$

$$\frac{b}{c} = 3.28; \quad P[F_{1.14} = 3.28] > .05$$

Source: Tables 44 and 48.

Table 50 shows the analysis of variance with the two criteria of classification. From this table we see that the regional differences are significant at the .001 level when differences in sex ratio classes are taken into consideration. The differences in sex ratio classes are still not significant even after taking regional differences into consideration, but the proba-

bility of $F_{1,14} = 3.28$ is close to .05. Although significance is not shown in this case, we see how it is frequently possible to show significance in one variable only after we take another variable into consideration.

Test for interaction. It is possible to make an additional test in a two-way classification when we have more than one case in each subclass. Since our illustration has three cases in each subclass, we will demonstrate the test. This test is called a test for interaction, and it is made to determine if there is interaction between the two criteria of classification. If there is significant interaction between the two, then we can use this fact to reduce our estimate of unexplained variation and, thus, frequently improve the level of significance or show significance where it could not be shown before. This amounts to considering an additional source of variation.

Testing for interaction in our case involves determining whether or not differences in sex ratio have the same effect on the percent of females single in one region that they have on the percent of females single in other regions. Stating in another way, it involves testing whether or not the regional differences are the same for the two sex ratio groups.

We make the test for interaction by first dividing our total variation into the variation between subclasses and the variation within subclasses. From Table 49 we see that we have six subclasses. By treating each of these subclasses as a class, we can get the sum of squares between subclasses. Subtracting the sum of squares between subclasses from the total sum of squares gives us the within-subclass variation. These computations are shown in Table 51.

The between-subclass variation of 159.33 includes the variation due to regions as well as the variation due to sex ratio, since these variations are computed from groupings of the subclasses. However, these two sources of variation do not explain all the between-subclass variation. The amount of between-subclass variation unexplained by the two criteria of classification is the interaction. Table 51 shows the sum of squares for interaction computed by subtracting sums of squares for regions and sex ratio classes from the between subclass variation. This gives an interaction sum of squares of 4.00.

The degrees of freedom for interaction are the product of the degrees of freedom of each of the criteria of classification, in this case, $(3 - 1)$ $(2 - 1)$ or 2. To test the significance of the interaction we form an F of the mean square variance for interaction and the mean square variance of the within-subclass variation. Table 52 shows the total analysis of variance and the $F_{2,12} = .33$ for testing the significance of the interaction. Since the F is less than one, it is clearly not significantly greater than the unexplained variance (within-class variance). Since the interaction is not significant, the analysis of variance made in Table 50 stands. If the inter-

Table 51. COMPUTATION FOR VARIATION DUE TO INTERACTION BETWEEN REGIONS AND SEX RATIO CLASSES IN PERCENTAGE OF FEMALES SINGLE IN 18 STANDARD METROPOLITAN AREAS

	Regions			Sums
	Northeast	Southeast	Middle States	
Sex ratios less than 94..........	21	20	19	
	26	19	19	
	27	18	23	
Sums....................	74	57	61	192
Sex ratios 94 and over..........	19	14	19	
	25	15	18	
	26	18	20	
Sums.................	70	47	57	174
Totals.....................	144	104	118	366

Total variation $= \Sigma X^2 - \dfrac{(X..)^2}{N} = 7{,}674 - \dfrac{(366)^2}{18} = 7{,}674 - 7442 = 232$

Between-sub-class variation $= \dfrac{(74)^2 + (70)^2 + (57)^2 + (47)^2 + (61)^2 + (57)^2}{3} - \dfrac{(366)^2}{18}$

$\qquad\qquad\qquad = 7{,}601.33 - 7{,}442 = 159.33$

Within-sub-class variation $= 232 - 159.33 = 72.67$

Interaction $=$ Between-subclass variation $-$ (Between-region variation $+$ Between-sex-ratio-class variation)

$\qquad\qquad = 159.33 - (137.33 + 18) = 4.00$

Source: Tables 41, 44, 48, and 49.

Table 52. ANALYSIS OF VARIANCE OF DATA OF TABLE 51

Source of variation	Sum of squares	Degrees of freedom	Mean square variance
Total.............................	232	17	
Between subclass....................	159.33	5	31.87
Between region....................	137.33	2	68.67
Between-sex-ratio classes...........	18.00	1	18.00
Interaction......................	4.00	2	2.00
Within subclass.....................	72.67	12	6.06

For testing interaction, $F_{2.12} = \dfrac{2.00}{6.06} = .33$

Source: Table 51.

action had been significant, then we would have recomputed the analysis of variance on the basis of Table 52 using the within-class mean square variance as the denominator of the F's.

Interpretation.   As a result of our analysis of this data on percentage of females 14 years of age and over single in 18 standard metropolitan areas, we can conclude that there is a real difference between the percentage of females single in metropolitan areas among these three regions. We have some evidence that the higher the sex ratio in these metropolitan areas the smaller the percentage of females single. This would seem to be a logical relationship, but the relationship we have observed could be the result of chance variations. On the basis of our findings we might be tempted to advise single women in metropolitan areas of the Northeast that they could improve their chances of marriage by moving to metropolitan areas of the Southeast. However, before we could give this advice with any degree of assurance, we would have to have additional information on age at marriage, marital status by place of marriage cross-classified by place of birth, etc.

Conditions for applying analysis of variance with two criteria of classification.   The first condition for the application of analysis of variance is the requirement that the sampling units be random samples from a universe to which we wish to generalize by means of an F test. This condition was approximately met in our illustration. With the exception of the Southeast the metropolitan areas used were randomly selected from the standard metropolitan areas in the regions involved. In the Southeast the standard metropolitan areas used were all those over 250,000 population in the region for which data were available at the time of writing.

Several other conditions of importance also exist. The most important of these perhaps is the requirement of approximately equal variance within the classes (homoscedasticity). We must also assume that the deviations from subclass means are uncorrelated and are jointly distributed in a multivariate normal distribution. In order to emphasize the main point of this discussion, we shall assume that these conditions have been met.

The important question becomes, "Are the methods of analysis of variance (and covariance) applicable for observational situations where data are observed in the cross-classifications in which they are found rather than in cross-classifications in which they have been randomly placed in an experiment?" The answering of this question in detail would require additional exploratory research as stated earlier. Only one minor suggestion will be offered. If exploratory analysis of observational data give results such as those of Table 50, which check with expectation based on theory and other sorts of analyses, it is possible that an empirical case for the use of such methods can be made.

ANALYSIS OF VARIANCE WITH ONE CRITERION OF CLASSIFICATION AND CLASSES OF UNEQUAL SIZE

The next analysis of variance will illustrate two procedures not yet introduced: (1) the use of class intervals of another quantitative characteristic as the classification; (2) the computation of the "between-class" sum of squares when the classes contain different numbers of items.

The problem. Another factor that might well affect the percentage of females 14 years of age and over that are single, is the size of the metropolitan area. The summary of the first analysis of variance, page 395, showed that the percentage of females single in the Northeast region was not significantly different from the percentage of females single in the Middle States. On this basis it was decided to take all the standard metropolitan areas with more than 250,000 population in the Northeast and Middle States for which data were available, classify them by population size, and examine the variation in percentage of females single in these groups.

The standard metropolitan areas were classified into three size groups: (1) metropolitan areas with populations between 250,000 and 499,999; (2) metropolitan areas with populations between 500,000 and 999,999; (3) metropolitan areas with populations over 1,000,000. Table 53 shows the percentage of females 14 years of age and over single in the standard metropolitan areas in each of these three size classes in the Northeast and Middle States.

Sum of squares. To perform an analysis of variance on the data of Table 53, we first compute the total sum of squares in the same way we would regardless of our scheme of classification. Next, we compute the between-class sum of squares by squaring each column sum, dividing by the number of items in that column (different for each column), summing the quotients for all columns, and subtracting the correction term, $\frac{(X..)^2}{N}$. In terms of the symbols of the computation guide on page 388 this is

$$\begin{matrix} \textit{Between-class} \\ \textit{sum of squares} \end{matrix} = \frac{(X_{.1})^2}{k_1} + \frac{(X_{.2})^2}{k_2} + \cdots + \frac{(X_{.m})^2}{k_m} - \frac{(X..)^2}{N} \qquad (8)$$

where k_i is the number of cases in the ith column.

The within-class sum of squares and the degrees of freedom are computed as before.

Analysis of variance. Table 54 shows the analysis of variance of Table 53. The analysis reveals that there are no significant differences in the proportions of females single in metropolitan areas among these three size

Table 53. COMPUTATION OF SUMS OF SQUARES FROM PERCENT OF FEMALES
SINGLE IN 31 STANDARD METROPOLITAN AREAS GROUPED BY SIZE

250,000–499,999 population	500,000–999,999 population	Over 1,000,000 population
19	22	24
28	19	24
25	19	21
25	26	22
20	16	19
19	23	23
23	21	27
24		18
26		20
27		25
18		
21		
20		
22		
Sums: 317	146	223
k_i 14	7	10

$N = 31$ $X.. = 686$ $\Sigma X^2 = 15,468$

Total sum
of squares
$= 15,468 - \dfrac{(686)^2}{31} = 15,468 - 15,180 = 288$

Between-class
sum of squares
$= \dfrac{(317)^2}{14} + \dfrac{(146)^2}{7} + \dfrac{(223)^2}{10} - \dfrac{(686)^2}{31}$

$= 15,195.83 - 15,180 = 15.83$

Within-class
sum of squares $= 288 - 15.83 = 272.17$

Source: 1950 Census of Population. Preliminary Reports. Series PC-5.

Table 54. ANALYSIS OF VARIANCE OF PERCENTAGE OF FEMALES SINGLE BY
SIZE CLASSES OF METROPOLITAN AREAS

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total......................	288	30		
Between-class	15.83	2	7.91	
				.81
Within-class	272.17	28	9.72	

$P[F_{2.28} = .81] > .05$

Source: Table 53.

classes of metropolitan areas. It might well be that if we had a larger size range, a range that included smaller population areas as well as these large ones, we might find a relationship between sizes of metropolitan areas and the percentage of females single. In the next chapter we will take up methods that will enable us to treat size as a quantitative variable rather than making groupings and treating it as a nonquantitative variable.

SUGGESTED READINGS

Dixon, Wilfrid J., and Massey, Jr., Frank J., *Introduction to Statistical Analysis* (New York: McGraw-Hill, 1951), Chap. 10.

Fisher, R. A., *Statistical Methods for Research Workers*, 10th ed. (London: Oliver and Boyd, 1948), Chaps. 7 and 8.

Johnson, Palmer O., *Statistical Methods in Research* (New York: Prentice-Hall, 1949), Chaps. 10 and 11.

Lindquist, E. F., *Statistical Analysis in Educational Research* (Boston: Houghton, 1940), Chap. 5.

McNemar, Quinn, *Psychological Statistics* (New York: Wiley, 1949), Chaps. 13 and 14.

Snedecor, George W., *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), Chaps. 10 and 11.

Total Correlation and Regression

Meaning of Correlation. The word "correlation" is used widely with a variety of meanings, both technical and nontechnical. Its general meaning is obvious from its etymology—the co-relation or the mutual relation between two or more phenomena. As explained in Chapter 20, statistics can treat only those problems of relationship where there are data on two or more *series* of phenomena. Hence, in statistics the first restriction imposed on the most general meaning of correlation is to limit its use to the problem of average relationship between two classes of phenomena with more than one item in each class. With such a restriction correlation still has a broad enough meaning to cover all the problems and methods of statistics of relationship. Therefore, to be able to differentiate between the specific bodies of method comprising statistics of relationship (listed on p. 348), we are imposing a further restriction. In this chapter we shall use correlation to denote the situations, along with the methods for their analysis, involving association between two or more classes of phenomena, each class of which is a series of measures on a quantitative characteristic.

Types of correlation and regression. Even the restricted meaning of the term correlation which we are adopting covers several types. There are two main twofold divisions of situations and their appropriate methods. The division by which we are separating the methods into chapters is (a) total correlation and regression, which analyze and describe the associations between two quantitative characteristics; and (b) multiple and partial correlation and regression, which analyze and describe the associations between three or more quantitative characteristics simultaneously. Within each of these types we shall consider the second division into linear and nonlinear correlation and regression.

The adjective "simple" is sometimes used to refer to linear correlation in contradistinction to nonlinear, sometimes to refer to total in contradistinction to multiple and partial, and sometimes to refer to total linear in contradistinction to the other three types—total nonlinear, multiple and partial linear, and multiple and partial nonlinear. Since the designation

"simple" is somewhat ambiguous, we shall not use it, but we shall designate types of correlation as divided in the preceding paragraph. In this chapter we shall deal with methods for studying relationships between two characteristics—that is, with the methods of total correlation and regression—considering first linear and then nonlinear total correlation and regression.

Linear Correlation and Regression: Method for Ungrouped Data. We shall present the methods of total correlation and regression, linear and nonlinear, by means of an example of association between two characteristics of economic areas. The data have all the questionable features described in the last section of Chapter 20.

The problem. There has been a growing interest in racial differentials in the Southeast and also in the level of living of farm operators. In our illustration we shall take the presence of running water in a farm dwelling as a partial index of level of living. We shall measure this as percent of farms reporting running water in the dwelling. The other measure is on the color of farm operators and is the number of nonwhite farm operators per 100 white farm operators. These two measures can be computed for state economic areas for 1945 from data in Donald J. Bogue's *State Economic Areas*.[1] The units on which these measures are taken for this illustration are the 31 economic areas in North Carolina, South Carolina, and Georgia. The question to be investigated is this—what are the facts about the association between nonwhite farm operators per 100 white farm operators and the percentage of farms reporting running water in the dwelling in 1945 in these 31 economic areas?[2] To make perfectly clear the mechanical set-up of a problem in correlation analysis, we can summarize the specifications of the example as follows:

SPECIFICATION OF A PROBLEM IN TOTAL LINEAR CORRELATION AND REGRESSION

Two quantitative characteristics whose relationship is being investigated:	1. Percentage of farms reporting running water in the dwelling (Y). 2. Number of nonwhite farm operators per 100 white farm operators (X).

---

[1] Donald J. Bogue, *State Economic Areas: A description of the procedure used in making a functional grouping of the counties of the United States* (Washington: Government Printing Office, 1951), Table B.

[2] Ideally we would investigate this relationship between running water in the dwelling and color of farm operators by the methods of contingency presented in Chapter 21. This would necessitate knowing for each farm whether or not it had running water in the dwelling and whether the farm operator was white or nonwhite. With this information we could make a 2x2 table and utilize the methods of contingency. However, since the data are not available in this form, we use correlation methods.

Series of units on which observations are available for varying degrees of incidence of the two characteristics:

The 31 economic areas of North Carolina, South Carolina, and Georgia.

Aspects of association to be investigated:

1. Existence of association.
2. Direction of association.
3. Degree of association.
4. Nature of association.

Reference of findings:

1. Findings from use of descriptive or historical methods to be referred to the association actually observed between the two characteristics as distributed among the 31 economic areas in 1945.
2. Findings from use of general or inductive methods to be referred to the association between the two characteristics in the infinite universe of possibilities from which this limited universe of 31 paired observations may be considered a random sample.

The data. For a problem in correlation and regression analysis the required data are measures on each of the two characteristics for a series of units. Table 55 shows a listing of measures on the two characteristics for the 31 economic areas. Let us emphasize that the data must be given in such a way that it shows for each unit its measures on *both* characteristics. The frequency distributions would not be satisfactory if given separately because the identity of each unit would be lost without a specification of the levels of incidence of both characteristics for it.

As with the description of single distributions, there are methods for the analysis and description of two distributions simultaneously for both grouped and ungrouped data. We shall consider first the methods for ungrouped data, using the data of Table 55, and later we shall consider the methods for grouped data using the data of Table 56.

Inspection of data. From inspection of Table 55 we see that the 31 areas range from 5.8 to 28.0 in the percent of farms reporting running water in the dwelling and from 1 to 195 nonwhite farm operators per 100 white farm operators. It is not easy to determine from inspection, however, much about the association between these two characteristics. Still one accustomed to such tables would be able to see that the areas having higher numbers of nonwhites per 100 whites on the average have a lower percentage reporting running water. This suggests an inverse or negative direction of association which can sometimes be detected from

Table 55. NUMBER OF NONWHITE FARM OPERATORS PER
100 WHITE FARM OPERATORS (X) AND PERCENT OF FARMS
REPORTING RUNNING WATER IN DWELLING (Y), ECO-
NOMIC AREAS OF NORTH CAROLINA, SOUTH CAROLINA,
AND GEORGIA, 1945

State economic area	Nonwhite farm operators per 100 white farm operators X	Percent of farms reporting running water in dwelling Y
North Carolina		
1 and A	1	28.0
2	4	16.8
3, B and C	27	17.5
4a	12	20.4
4b	13	27.4
5 and D	28	20.2
6 and E	46	9.7
7	125	5.8
8	72	7.9
9	106	9.0
10	39	8.5
11	48	7.8
South Carolina		
1	12	12.9
2	43	19.6
3	76	12.8
4	104	13.5
5 and A	62	17.3
6	180	10.6
7	93	7.3
8 and B	176	9.7
Georgia		
1 and A	8	14.3
2	1	18.3
3 and B	12	18.6
4a	65	12.5
4b	123	13.5
5 and C	75	20.6
6	73	9.6
7a	195	14.4
7b	69	15.2
8	32	13.9
9 and D	33	18.7

Source: Donald J. Bogue, *State Economic Areas: A Description of the
Procedure Used in Making a Functional Grouping of the Counties of the
United States* (Washington: Government Printing Office, 1951), Table B.

Table 56. CROSS TABULATION OF 31 ECONOMIC AREAS BY PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING AND NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS, 1945

Percent of farms reporting running water in dwelling	All numbers	Number of nonwhite farm operators per 100 white farm operators						
		0–24	25–49	50–74	75–99	100–124	125–149	150–174	175–199
All percentages	31	8	8	5	3	3	1	0	3
27.5–29.9	1	1							
25.0–27.4	1	1							
22.5–24.9	0								
20.0–22.4	3	1	1		1				
17.5–19.9	5	2	3						
15.0–17.4	8	1		2					
12.5–14.9	8	2	1	1	1	2			1
10.0–12.4	1								1
7.5– 9.9	7		3	2		1			1
5.0– 7.4	2				1		1		

Source: Table 55.

the mere listing of measures. If the series of units were much longer, however, it would be more difficult to detect any association. Or if there were only a low degree of association, it would be difficult to detect. At any rate, the information one can gain from inspection of ungrouped data in this form is too vague and is neither condensed nor precise enough to be conveyed to another satisfactorily.

Graphic analysis. The construction of a scatter plot or scatter diagram is usually the first step in correlation analysis. In Figure 38 a dot is made for each area at a point determined by using one measure for that state as the X coordinate and the other measure for that area as the Y coordinate. The making of a scatter plot is not an essential step in correlation analysis—that is, one may determine the coefficients of correlation and regression without first making a scatter plot—but it is advised as a first step for several reasons. The reason which is important at this stage of the analysis is that the scatter plot quickly gives the person who is accustomed to using it a rough idea as to all four aspects of association. In the present case one can see there is association because the pattern of the dots is such that its axis in one direction (upper left to lower right in this example) is longer than its axis perpendicular to the first; that the direction of association is negative because the dots toward the left side of the plot are on the average higher than those on the right, meaning that values higher than the mean in the percentage reporting

Figure 38.   Scatter Plot of Percent of Farms Reporting Running Water in Dwelling (Y) and Number of Nonwhite Farm Operators per 100 White Farm Operators (X), 31 Economic Areas, 1945.

running water tend to occur with values lower than the mean in the number of nonwhites per 100 whites; that the degree of association is what we term "moderate" rather than "high" because the dots do not cluster very closely around the longer axis; that the nature of association is such that on the average the percent of farms reporting running water in the dwelling decreases 5 or 6 percent for 100 units increase in number of

nonwhites per 100 whites. Such a rough description of association as is afforded by a scatter plot often suffices to let the investigator know whether or not to continue a more detailed correlation analysis which will be time consuming. If in the present case the scatter plot had shown there was practically no association, it would have been unnecessary to pursue the analysis further.

The correlation coefficient. It is obvious that the description of association between the two characteristics afforded by the scatter plot is still not so precise as one might wish. Therefore, we turn to a summarizing measure, the coefficient of correlation, which summarizes in one coefficient a description of the existence, direction, and degree of association. There are many way of defining the coefficient of correlation, but there is no satisfactory definition which does not involve statistical concepts foreign to everyday thinking. Therefore, the student must try to translate the statistical concepts into something which corresponds to everyday thinking in his own mind.

Let us review the meaning of the concept of "total variation" or simply "variation," as defined in Chapter 9. The variation of a distribution is equal to the sum of the squares of the deviation of each measure from the mean of the distribution. The definition of variation in terms of symbols and the actual formulas by which we compute the variation of a distribution are as follows:

$$\text{Variation} = \Sigma(X - \bar{X})^2 = \Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \tag{1}$$

When we are considering two distributions simultaneously in the study of their relationship, we must differentiate between them by assigning them different letters. If there are only two distributions, we usually use the letters X and Y, letting X refer to the one we regard as "independent" or in some way as being primary in time or in any causal sequence. Statistical methods cannot help one in deciding which variable should be considered as "cause" and labeled X, and which as "effect" and labeled Y. In fact, there may be no causal relationship at all between the two variables studied, or there may be a mutually interacting relationship, in which case it makes no difference which one is called X and which Y. Usually, however, the nature of the problem indicates which one we are investigating as a possible condition from which we may predict the level of the other. In this particular problem we are concerned more with the ratio of whites to nonwhites as a condition from which we may predict the proportion of farms reporting running water in the dwelling rather than vice versa. We shall designate the number of nonwhite operators per 100 white farm operators as X and the percentage of farms reporting

running water in the dwelling as Y. The quantity designated by formula (1) above represents the variation of the 31 economic areas in their number of nonwhite farm operators per 100 white farm operators, while the variation in the percentage of farms reporting running water is

$$\Sigma(Y - \bar{Y})^2 = \Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} \qquad (2)$$

It is hoped by now that the student has had enough familarity with the concept of variation as here defined to feel at home with it, to imagine it perhaps visually or in some other way as a quantity built up by a contribution from each measure in a distribution—the square of its distance from the mean—with many smaller contributions from the numerous measures clustering around the mean and with a few large contributions from the extreme measures at or near the ends of the observed range. For one must learn to think in terms of these individual deviations, their squares, and the sum of their squares (variation) to grasp the meaning of a correlation coefficient, since correlation coefficients are defined in terms of the variation of the two distributions and a new concept which we shall now introduce, their *covariation*.

If for each unit on which observations of the degree of incidence of the two characteristics have been made we form the product of the deviation of its measure in one distribution from the mean of that distribution and the deviation of its measure in the other distribution from the mean of that distribution, and if we sum these deviation products for all units, we get a quantity called the *covariation* of the two distributions. The definition in symbols is

$$Covariation = \Sigma(X - \bar{X})(Y - \bar{Y}) = \Sigma xy$$
$$= \Sigma X Y - \frac{(\Sigma X)(\Sigma Y)}{N} \qquad (3)$$

Now we are able to define the coefficient of correlation. It is the ratio of the covariation to the geometric mean [3] of the variations of the two distributions.

$$Coefficient\ of\ correlation = r_{YX} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \qquad (4)$$

Without puzzling too much at the moment over the meaning of r (usually written without subscripts if only two variables are being considered), let us go ahead and compute it. To evaluate formula (4) we shall need only the three quantities, Σx^2, Σy^2, Σxy. The student has already learned how to compute the first two for ungrouped data, and needs to

---

[3] The geometric mean of two quantities is the square root of their product.

learn only one new process to compute the third. If a calculating machine is available which accumulates squares or products, the process is not long. From the original measures of Table 55 one secures the following basic data for the computation of the correlation coefficient:

$$\Sigma X = 1,953 \qquad \Sigma Y = 452.3 \qquad N = 31$$
$$\Sigma X^2 = 209,699 \qquad \Sigma Y^2 = 7,529.03 \qquad \Sigma X Y = 23,868.7$$

The only new expression here is $\Sigma X Y$. It is obtained by multiplying for each state its two measures and adding these products.

The next step is the computation of the quantities Σx^2, Σy^2, Σxy from the evaluation of the formulas (1), (2) and (3).

$$\Sigma x^2 = 209,699 - \frac{(1953)^2}{31} = 86,660$$

$$\Sigma y^2 = 7,529 - \frac{(452.3)^2}{31} = 929.828$$

$$\Sigma xy = 23,868.7 - \frac{(1953)(452.3)}{31} = -4,626.2$$

The final step is to compute r by substituting these values into (4).

$$r = \frac{-4,626.2}{\sqrt{(86,660)(929.828)}} = -.515$$

An alternate procedure is to compute r from the original sums by evaluating the formula,

$$r = \frac{N\Sigma X Y - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad (5)$$

$$= \frac{31(23,868.7) - (1953)(452.3)}{\sqrt{[31(209,699) - (1953)^2][31(7529.03) - (452.3)^2]}}$$

$$= -.515$$

Interpretation of correlation coefficient. Let us see what information the coefficient of correlation yields about the association between the percentage of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators. We are considering the coefficient now in only its descriptive function, that is, in describing the aspects of the association between these characteristics for the 31 economic areas observed. The fact that r is different from zero answers the question as to *existence* of association in the affirmative. The fact that r is negative indicates that the *direction* is inverse—that is, areas with higher percentages reporting running water in the dwelling

have on the average lower numbers of nonwhite farm operators per 100 white farm operators. The absolute value of r, .515, indicates that the *degree* of association is "moderate." "Moderate" is a subjective sort of word, but it will suffice to suggest a level until experience and other methods of analysis help one to learn what amount of importance to attach to a coefficient of correlation of this size. Thus, we see that the coefficient of correlation describes three of the four aspects of relationship but that it does not give any information on the *nature* of the relationship between the two characteristics (as we have defined nature to mean the differences or changes in one measure associated with unit changes in the other measure). Description of the nature of association requires the computation of another coefficient.

The coefficient of regression and the regression equation.[4] The nature of the association between two quantitative characteristics can be described by means of an algebraic equation called the regression equation, the estimating equation, or the predicting equation. Let us assume that this equation will be of the form,

$$Y_e = a + bX \tag{6}$$

where Y_e will be the estimated or predicted value of the characteristic designated by Y for any corresponding value we may wish to assign to X. The assumption underlying the use of this form of an equation to describe the nature of the association between two characteristics is this—that equal differences in the measures of one characteristic are associated with equal differences in the measures of the other characteristic. Expressed in algebraic terms, this assumption is that the relation between the two characteristics is *linear*, meaning that it can be described in an equation of the first degree,[5] which when plotted gives a straight line. This assumption is not always justified for the association between two characteristics, but we cannot learn how to test it precisely until the methods of curvilinear correlation have been presented. One can test it roughly, however, by looking at the scatter plot of Figure 38 and seeing that the dots tend to cluster around a straight line drawn from the upper left portion of the plot to the lower right portion. Because the dots scatter considerably from any straight line, we cannot be certain that a curved line would not fit better, but we shall assume linearity until we can apply a more rigorous test.

The problem of describing the nature of the association then becomes

---

[4] The procedures of this section are the same as those presented for fitting a straight line to describe a trend in a time series. However, we shall proceed as if they had not been presented since some may wish to study correlation before time series.

[5] A first degree equation has only the first powers of X and Y occurring, these only in the numerators of terms, and with no product terms involving both X and Y.

that of finding what values of a and b will give the line which "best fits" the dots of the scatter plot. Various criteria might be used to determine what is the "best" fitting line, but the one commonly employed is the criterion of "least squares." The line which fulfills the criterion of least squares is that line from which the sum of the squared deviations of all the dots is a minimum. If we designate the observed measures of the Y characteristic as Y's and the Y coordinates of the points of the line which have the same X values as the observed values as Y_c's, then the least squares criterion can be expressed in a combination of symbols and words thus,

$$\Sigma(Y - Y_c)^2 \text{ must be a minimum.} \qquad (7)$$

This means that for any line other than the one we shall determine, the sum of the squares of the deviations of the observed Y's from their corresponding Y_c's will be greater than the sum from the best fitting line.

It is immediately apparent that there is a parallelism between the least squares regression line and the mean of a single distribution, from which the sum of the squares of the deviations of the measures is also a minimum. This regression line may be thought of as an extension of the mean when it is projected into another dimension—that is, when another distribution is being considered simultaneously. For any value of the second characteristic it represents the mean value of the first characteristic expected, if the relation between them is linear.

Computation of a and b. It is possible to find for any two distributions the value of a and b for equation (6) from the same data used in computing the coefficient of correlation. Let us note that the regression equation, unlike the correlation coefficient, is not symmetrical with respect to X and Y. Equation (6) may be written more fully,

$$Y_c = a_{YX} + b_{YX} X \qquad (6A)$$

with the subscripts denoting that the coefficients belong to the equation for the regression of "Y on X," which means the equation for computing or estimating values of the Y characteristic when values of the X characteristic are known. This is the customary form of the regression equation since we customarily estimate or predict the "dependent" characteristic Y from the "independent" characteristic X. Sometimes, however, we have occasion to do the reverse, in which case we write the regression equation in the form,

$$X_c = a_{XY} + b_{XY} Y \qquad (6B)$$

and compute values for a_{XY} and b_{XY}. When the coefficients a and b are used without subscripts, however, they refer to equation (6A) not to (6B).

The formula for computing b is

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} \tag{8}$$

and that for computing a is,

$$a_{YX} = \frac{\Sigma Y - b\Sigma X}{N} \tag{9}$$

Using data from page 413, we make the following computations for the regression coefficients of the regression equation of the percent of farms reporting running water in the dwelling (Y) and the number of nonwhite farm operators per 100 white farm operators (X), in 1945 for 31 economic areas.

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{-4626.2}{86,660} = -.05338$$

$$a = \frac{\Sigma Y - b\Sigma X}{N} = \frac{452.3 - (-.05338)(1953)}{31} = 17.95$$

An alternate procedure is to use a formula for b which requires only the original sums from page 413 rather than the intermediate measures, thus

$$b = \frac{N\Sigma X Y - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \tag{8A}$$

$$b = \frac{(31)(23,868.7) - (1,953)(452.3)}{(31)(209,699) - (1953)^2} = -.05338$$

We can now write the regression equation for our example by substituting these numerical values computed for a and b in equation (6), thus,

$$Y_c = 17.95 - .05338\,X \tag{10}$$

Equation (10) is a concise mathematical description of the nature of the association manifested between the two variables. It is necessary to learn to translate the equation into less mathematical terms for the non-mathematically inclined.

Graphic representation of regression equation. The principles of analytic geometry make it possible to translate an algebraic description of the relation between variables into a graphic description. If the algebraic description is an equation of this form, its graphic equivalent will be a straight line. To plot a straight line we need to know only two points. To plot two points we need to know two pairs of corresponding values of X and Y_c. These values may be obtained by choosing any two values of X

within the range represented on the scatter plot and substituting them, one at a time, in the regression equation (10). Let us choose for values of X, 0 and 100. Then for the first point, when $X = 0$, $Y_c = 17.95$. For the second point, when $X = 100$, $Y_c = 17.95 - (.05338)(100) = 12.612$. In Figure 39 these two points, designated by their coordinates, (0, 19.75) and (100, 12.612) are shown as circles on the scatter plot. When these two points are plotted, we draw a straight line through them and this line is the geometric form of the regression equation (10). It is called the regression line or the line of regression of Y on X.

For any point one might choose on the regression line, he can read off corresponding values of X and Y_c. For instance, the point where the line crosses the vertical line rising from the value 50 on the X scale has a corresponding Y_c value of about 15 read from its height on the Y scale. For any point on the line the Y_c value means the percent of farms reporting running water in the dwelling that one would expect a subregion to have if the subregion had the corresponding X value as its number of nonwhite farm operators per 100 white farm operators, and *if* the association between the two characteristics were perfect—that is, if all the variation between the subregions in the percent of farms reporting running water in the dwelling were "explained" or "accounted for" by the variation in the number of nonwhite farm operators per 100 white farm operators.

The scatter of the dots about the line shows that the association between these characteristics is not perfect. This is reasonable, for we know that factors other than ratio of white to nonwhite farm operators influence the presence of running water in farm dwellings. We shall learn to measure the amount of scatter presently. We may observe now, however, that the inverse of scatter, the closeness with which the dots cluster about the line, is measured by the correlation coefficient.

Let us note how the values of a and b of the regression equation are represented in Figure 39. We saw that in equation (10) when $X = 0$, the second term on the right hand side of the equation vanishes and the corresponding value of Y_c is 17.95, or a. In Figure 39 the point where the regression line cuts the Y axis is 17.95 units above zero on the Y-axis. This distance is called the Y intercept and is always equal to the value of a. The value of a is not of special importance for further analysis, and often, when the X variable cannot take a zero or negative value, it is meaningless; but geometrically, at least, it can be interpreted to mean the value of Y_c when $X = 0$.

The meaning of b is of more importance although it is not quite so easily read from a chart. The value of b describes what is called the "slope" of the regression line, which is the number of units change in the height of the regression line for each unit increase in X. In our example

Figure 39. Regression of Percent of Farms Reporting Running Water in Dwelling (*Y*) on Number of Nonwhite Farm Operators per 100 White Farm Operators (*X*), 31 Economic Areas, 1945. (Source: Table 55.)

this means that if the regression line described the relation between *X* and *Y* perfectly, then for each increase of one nonwhite farm operator per 100 white farm operators the percent of farms reporting running water in the dwelling would change by −.05338, or decrease by about five one hundreths of a percent.

The coefficient *b* is often called "the" regression coefficient, although *a* is likewise a coefficient of the regression equation. The sign of *b* like the sign of *r* is determined by the sign of Σ*xy*, the numerator of each, since

the denominator of each is always positive. Thus, b as well as r describes the direction of the association.

As we have done with other bodies of method of analyzing association, let us summarize the description of the association of the two characteristics as it existed in the 31 economic areas in 1945.

SUMMARY OF DESCRIPTION OF ASSOCIATION BETWEEN PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING AND NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATIORS IN 31 ECONOMIC AREAS, 1945

Aspect of association	Description
1. *Existence of association*	The two characteristics are associated ($r \neq 0$).
2. *Direction of association*	The two characteristics are inversely associated (r is negative).
3. *Degree of association*	The degree of association between the two characteristics is "moderate," and is measured precisely by the absolute value of r, .515.
4. *Nature of association*	(Under the assumption of linearity) the nature of the association is described by the regression equation, $Y_c = 17.95 - .05338X$.

DESCRIPTION OF ASSOCIATION FOR THE UNIVERSE

Generalization of the above results. The observations of the measures of the 31 economic areas of North Carolina, South Carolina, and Georgia on the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators may be considered as comprising an entire limited universe. There is no real sampling situation here because these states cannot be considered as a sample from which we can generalize to any other area. Nevertheless, even in research articles which include correlation coefficients based upon measures of the 48 states as varying units, one usually finds tests of significance, or standard errors, or other devices which have been developed from sampling theory. It is again the situation where one imagines a hypothetical infinite universe of all the possible limited universes of 31 pairs of observations which might have been observed. This hypothetical universe is the universe from which the 31 pairs of observations analyzed may be considered a random sample. We refer the readers to the chapters of Part III for a fuller discussion of the concept of the universe of possibilities: we merely suggest here its meaning in terms of this one problem. Two people might discuss the problem thusly:

A. I have finished the description of the association between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators in 1945 as summarized above. The description is complete and exact for the 31 economic areas of North Carolina, South Carolina, and Georgia. There is no need to make any tests of significance of the coefficients determined.

B. But are you not interested in the relationship between the two characteristics in general rather than just in the way they were associated in this particular situation?

A. I would be interested in their association "in general" if I knew precisely what "in general" means for these characteristics. I certainly couldn't generalize these results to other decades where the factors might have been behaving differently; and certainly not to the entire United States or to other countries where the factors might have been behaving differently. My information is only on the 31 economic areas of North Carolina, South Carolina, and Georgia for 1945, and I cannot project the description of an association of characteristics beyond the time and place to which the basic information refers.

B. I grant that the description would not be valid for other times or other places except insofar as other relevant conditions in them are the same as in this actual situation. But for the time and place to which these data relate, can you not from this analysis and description of the association between the two characteristics generalize beyond the historical account of the unique facts recorded?

A. What is there beyond a historical account? What manifestations of association between the two characteristics could there possibly have been other than these which actually happened and were recorded as the set of observations I analyzed?

B. This is just the question. Given the same conditions which produced the running water in the dwelling and the ratio of white to nonwhite measures, operating repeatedly in the same place and time, would they always produce identically the same results?

A. But your question refers to a hypothetical situation, for we cannot imagine 1945 being repeated with all factors influencing these two measures remaining the same.

B. If, for the moment, you do imagine such a hypothetical situation, what would you expect as the resulting association between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators?

A. I would expect minor variation in the coefficients describing association, but essentially the same sort of association that I have described.

B. Do you have any idea of the magnitude of the variation which might be expected to occur?

A. Only this idea, drawn from analogy, that it would probably be the type of variation observed in physical experiments when all the factors known to influence a result are held constant, but still the results show a residual variation.

B. The type of variation you are describing, which is produced by a very

great number of very small factors, is called "random" variation, and a great deal of theoretical work has been done in describing the form and amount of random variation to be expected under various assumptions.

A. But since I *know* what happened in these 31 economic areas in 1945, what difference would the existence of random variation in a hypothetical situation make?

B. Suppose that in the imaginary universe of all the possible sets of observations on running water in farm dwellings and the ratio of white to nonwhite farm operators which might have occurred during 1945 there is really no correlation between the two characteristics. Then if correlation coefficients were computed for every set of 31 observations, their mean would be zero. Most of these coefficients would cluster around zero in value with a small proportion of them having moderately great positive or negative values. Suppose in such a case you have drawn in your particular set of 31 observations a value of the coefficient which is no greater than one would expect fairly frequently just from random variation. Would not that fact have a bearing on your interpretation, in terms of your problem, of the description of association you have made?

A. Why yes, if the association I have described could be regarded as "accidental" in the sense of being fairly common in a sample of 31 drawn from a universe with no association, then I would not attach much importance to it. Doing so would give the false impression that the two characteristics were meaningfully associated in some way—whether by an intermediate factor or by some common antecedent factor or factors I do not know—when the association is possibly "accidental," if "accidental" can be defined in this way.

B. Then wouldn't it make you feel more secure in using your results for interpreting the situation, or for planning further research, if in some way you found that an association of the degree you observed would *not* be expected to occur one time out of 100 in samples of 31 from a universe with no correlation?

A. I suppose so, for in spite of my carefully qualified statements of interpretation, a correlation coefficient does make me suspect an actual relationship between the two characteristics, be it ever so indirect. And I should want to know that I was not wasting my time looking for the meaning of something which might be regarded as "accidental." Nor would I want to suggest to those who read my report that there is a meaningful relation between the two characteristics if the only evidence I have may be regarded as "accidental."

B. Well, that is exactly what tests of significance of correlation coefficients do. They tell you what is the probability that a coefficient of the size you find for the number of observations you have might occur "accidentally."

A. But does "accidentally" as you have defined it actually apply to the situation we are discussing?

B. That is the crucial point of the whole argument as to whether or not you should use tests of significance in such a situation as this. One might justify the use of the concept of "accidental" in either of two ways. The first justification is based upon the assumption that demographic characteristics, like others that can be tested empirically, are subject to a residual fluctuation of the "random" variety when all the known influencing factors remain the same, and that any magnitude of fluctuation which is expected fairly frequently

from these unknown but probably innumerable and minor factors, is called "accidental." Since the assumption can never be tested empirically—we cannot, for instance, turn time back and repeat 1945—there can be no proof as to whether this assumption is justified.

The second justification is more abstract. Without any analogy to physical experiments or to any real situation, "accidental" is simply defined as encompassing the range of fairly frequently observed fluctuations which can be theoretically deduced for a random sampling situation. Then a hypothetical universe is postulated, a universe from which the actual set of observations may be considered a random sample, and the information obtained from the actual observations is used to infer information about this hypothetical universe. The universe and the tests of significance relating to it are simply logical structures. They do not have any objective counterparts; they have no everyday interpretation. There is no one convincing argument which justifies their use; but many statisticians feel that they may be useful in getting at a concept of the regularity or order underlying social phenomena.

A. Neither of your arguments convinces me of the utility of tests of significance where there is no practical sampling situation. Since I am merely beginning the study of statistics as applied to sociological research, however, I am willing to examine carefully the practice of persons with more experience in the field. What do other sociologists do about tests of significance in such situations?

B. They do everything imaginable. Some ignore tests of significance altogether, implying the same point of view which you hold. Others attach to their coefficients of correlation either their probable [6] or standard errors, assuming that their readers will know what to do with them. Others tabulate "critical ratios," (the ratio of an observed coefficient to the standard error of samples of the same size in a universe which has the same value of its coefficient as the observed), also assuming that their readers can make the correct interpretation of them. Others merely state whether or not an observed coefficient is "statistically significant" without specifying what tests they use—and incorrect tests as well as correct tests are in common use. Still others, though fewer, give the results of specified correct tests of significance. But even in the case of the latter, there is seldom an explanation of what they mean by labeling a particular coefficient of correlation as "significant" or "statistically significant." Therefore, you can find good precedent among sociologists for whatever you choose to do—except for explaining exactly what you mean when you use tests of significance.

A. Then if after more thought on the matter I decide to test the significance of my coefficients, in my report of a quantitative sociological research project should I go through all of the explanation you have given of the logical structure, the universe of possibilities, the nature of random variation, etc. every time

---

[6] The probable error, which assumes a symmetrical sampling distribution, cannot be applied correctly to any coefficient of correlation different from zero, for only in universes where there is no correlation is the sampling distribution of r's symmetrical. Therefore, the probable error should *never* be attached to an observed coefficient.

I give the results of the application of a test of significance to an observed correlation coefficient?

B. Probably not. Such a procedure would be repetitious and boring. One himself, however, should understand the meaning of his tests; he should choose correct tests; and at this point of development of the application of statistical methods to sociological research, he should specify the test he has used. But what is of equal importance is that the reader should know what the test means, since otherwise he may be unduly impressed with the magic phrase, "statistically significant." It is an example of the common situation where the research worker should use correctly and carefully the appropriate statistical methods, but because of the necessity of economy of words in reporting his results, he has to assume the statistical literacy of his readers.

Since tests of significance of correlation coefficients are found increasingly in sociological literature, regardless of whether those who take the point of view of B in the discussion above can *prove* one should use then (when there is no practical sampling situation), one should learn how they can be made correctly. The most common test of significance answers for the universe the question as to whether or not association *exists* in the universe—that is, it investigates for the universe the first aspect of association.

Existence of association in the universe. The problem of establishing the fact of the existence of association in the universe leads to the answering of this question for our example—in the infinite universe of possibilities of all sets of 31 observations on the percent of farms reporting running water in dwellings and the ratio of white to nonwhite farm operators which might have been made on these economic areas in 1945, is there a negative association? As before, we have to set up a null hypothesis and to show the improbability of observing what we did under such a hypothesis. We shall use the same sequence of steps we used in testing hypotheses about single distributions.

1. *Formulation of the hypothesis to be tested.* What we want to show is that a negative association between the percent of farms reporting running water in dwellings and the number of nonwhite farm operators per 100 white farm operators exists in the infinite universe from which our 31 observations may be considered a random sample. Using the Greek letter rho to represent the universe parameter corresponding to the statistic r, we may express as the hypothesis we wish to establish,

$$\rho < 0$$

The most general null hypothesis which includes all possibilities other than the one we wish to establish is

$$\rho \geq 0$$

The limiting case of this general null hypothesis, for which negative r's, such as we have observed, are most likely to be observed in samples, is the specific null hypothesis to be tested and may be formulated thus,

$$\rho = 0$$

2. *Description of the sampling distribution of r or $\hat{\rho}$.* If the number of observations on which it is based is not too small, the sampling distribution of the estimate of the universe coefficient of correlation, $\hat{\rho}$, made from samples of N drawn from a universe where the correlation is zero has a mean of zero, an approximately normal form, and a standard deviation of,

$$\sigma_r = \frac{1}{\sqrt{N-1}} \tag{11}\,[7]$$

Since the number of observations is greater than 30, we may use this formula for the standard error, *not* of the estimate of the parameter in a universe where $\rho = -.515$, but of an estimate of the parameter in a universe where $\rho = 0$ (in order to be consistent with the hypothesis being tested), and may refer our results to the normal curve. Substitution of $N = 31$ in (11) gives us

$$\sigma_r = \frac{1}{\sqrt{31-1}} = .1826$$

3. *Determination of the probability that a deviation as unusual as that observed would be expected under the hypothesis.* The deviation of the observed $r = -.515$ from the mean of the sampling distribution, $\rho = 0$, is

$$-.515 - 0 = -.515$$

The ratio of the deviation to the standard deviation of the sampling distribution is

$$\frac{-.515}{.1826} = 2.820$$

From Table C of the Appendix we find that the probability that a deviation as unusual as one of 2.82 standard deviation units from the mean would be observed in a normal distribution is .0048.

4. *Rejection of the null hypothesis.* Because the P of the above section is less than .01, we reject first the specific null hypothesis, $\rho = 0$, and then the more general null hypothesis, $\rho \geq 0$. The only alternative, which we are led to believe is the correct statement, is that $\rho < 0$, which means that

---

[7] Note that formula (11) is to be used *only* in testing the hypothesis that $\rho = 0$. This is *not* the general formula for the standard error of a coefficient of correlation drawn from a universe where ρ is different from zero.

the "true" or "real" coefficient in the universe of possibilities is negative and different from zero.

5. *Interpretation of the test in terms of the problem.* Since the observed coefficient of correlation is greater in absolute value than would be expected in 48 out of 10,000 times in samples of 31 from a universe with no correlation, we may conclude that the negative association observed between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators in "significant." This means that economic areas with a high ratio of nonwhite farm operators had on the average low percentages of farm dwellings reporting running water while those economic areas with a low ratio of nonwhite to white farm operators had higher percentages of farm dwellings reporting running water than could be termed "accidental." The assigning of priority in relationship to one or the other of the two characteristics is not possible by correlation analysis.

Direction of association. In the test made above for the existence of association in the universe the direction of association in the universe has been established with the same degree of confidence or at the same level of significance as the existence of association. We need merely to restate the findings of the above section with slightly different emphasis—that is, our findings justify the conclusion that the association between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators in 1945 is negative in the universe of possibilities.

Degree of association. The test for the existence of association in the universe has indicated that the two factors *are* definitely associated, but it has not indicated how *closely* they are associated. The distinction between the two aspects of association, existence and degree, must be grasped clearly, for the two aspects are often confused. As suggested in the discussion of A and B, the person who does not clearly understand the implied test of the hypothesis that $\rho = 0$, when he reads that a certain r is "significant," is likely to think "significant" is being used in its nonstatistical sense to mean "important," and a significant r does *not* necessarily mean an r of size great enough to be important. An observed r of .1 may be significant if based upon a great enough number of observations, yet it can hardly be called important since it means that only one percent of the variation in one characteristics is associated with the variation of the other characteristic. On the other hand, an observed r of .8 may not be significant, if it is based on a very few observations, and yet for the sample at least it is important, for it means that 64 percent of the variation in one characteristic is associated with variation in the other characteristic. (The method by which these percentages are determined will be given on pages 429–430.)

First, let us estimate the universe value of ρ from the observed r. As has been mentioned before, there is no unanimity of opinion on the criteria for "best" estimates of universe parameters, and the correlation coefficient is one of the measures for which different criteria yield different formulas for estimation. The "criterion of maximum likelihood" gives the following formula,[8]

$$\hat{\rho} = r = \frac{\Sigma xy}{\sqrt{\Sigma x^2\ \Sigma y^2}} \tag{12}$$

Substitution in this formula gives us an estimate of the universe value of the coefficient of correlation between the two characteristics,

$$\hat{\rho} = r = -.515$$

For an estimate we always wish if possible to have a measure of its reliability or precision. To derive such a measure we need to know the sampling distribution of estimates of $\hat{\rho}$ made from samples of size 31. In the case of the correlation coefficient, however, the size of the standard deviation of the sampling distribution is different for every different universe value of ρ. Formula (11) gave the value of the standard error (standard deviation of the sampling distribution) of r when $\rho = 0$. The more general formula for the standard error of a coefficient of correlation is

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{N - 1}} \tag{13}$$

Note that the formula for the standard error of an observed r calls for the value of the universe ρ, which is of course unknown in any practical situation. This presents one difficulty in proceeding to set up confidence limits as we did in the case of an estimate of the universe mean. A second difficulty arises from the fact that the form of the sampling distribution is not normal; in fact, it is not even symmetrical except when $\rho = 0$. In a skewed distribution equal deviations measured on either side of the mean do not cut off equal areas. It is because of the skewness of the sampling distribution of the coefficient of correlation that the concept of "probable error" is inappropriate for measuring the unreliability of an estimate of a universe correlation coefficient. (The probable error of a measure is the semi-interquartile deviation of the sampling distribution of that measure, *if* the sampling distribution is symmetrical, so that the deviation may be determined which cuts off 25 percent of the area on either side of the mean.)

---

[8] See L. H. C. Tippett, *The Methods of Statistics*, 2d ed. (London: Williams and Norgate, 1937), pp. 164–165. For another formula for making an estimate of the universe parameter see Mordecai Ezekiel, *Methods of Correlation Analysis* (New York: Wiley, 1941), p. 121.

The skewness of the sampling distribution is not marked unless ρ is rather high, approaching .8 or greater. For lower values of ρ the sampling distribution of r is nearly normal if N is not small, and approximate 99-percent confidence limits might be set up thus,

$$\hat{\rho}_{1 \text{ and } 2} = \hat{\rho} \pm 2.58\hat{\sigma}_r \tag{14}$$

using $\hat{\rho}$ for ρ in formula (13) and thus obtaining from it only an estimate of the standard error, $\hat{\sigma}_r$, rather than the standard error itself, σ_r. There is a more accurate procedure, however, which gets around both of the difficulties mentioned—that we do not know ρ and, therefore, cannot compute σ_r; and that the form of the sampling distribution of r is not normal.

The procedure is an invention of R. A. Fisher involving a transformation called the Z or z' transformation. (Fisher himself refers to it as the z transformation but this leads to confusion of the symbol with that for his z distribution used with analysis of variance.) It is the fact of the limited scale of r which makes its sampling distribution skewed and makes the confidence limits determined by formula (14) inaccurate. Let us consider, for example, the sampling distribution of r obtained from samples of size 17 drawn from a universe with a ρ of .9. The standard error of r from formula (13) is

$$\sigma_r = \frac{1 - (.9)^2}{\sqrt{17 - 1}} = \frac{1 - .81}{\sqrt{16}} = \frac{.19}{4} = .05 \text{ (approximately)}$$

Now suppose we had actually observed an r of .9 and tried to set up confidence limits by formula (14). The upper limit would be greater than 1.0, the maximum value r or $\hat{\rho}$ can take.

The transformation is designed to change the absolute value of r, which can vary only from zero to one, into a related measure Z, which can vary from zero to plus infinity. The actual transformation is described by the formula,

$$Z = \frac{1}{2} \log_e \frac{1 + r}{1 - r} \tag{15}$$

Table G in the Appendix gives values of Z corresponding to values of r, making actual substitution in formula (15) unnecessary. To the student unfamiliar with the mathematical custom of transforming variables, we can explain what we are going to do as follows. Our problem is to establish confidence limits for $\hat{\rho} = r$, but we can set up confidence limits by adding and subtracting a certain value to the estimate only for statistics with symmetrical sampling distributions, whose standard errors can be found. Since r is not symmetrically distributed, we shall transform our observed r into another related measure, Z, which *is* symmetrically and approximately normally distributed, set up confidence limits, Z_1 and Z_2, for

this measure, then re-transform the two confidence limits of Z back into corresponding values for r, that is into r_1 and r_2 and thus obtain the confidence limits of our estimate of $\hat{p}$. Aside from the advantage of being approximately normally distributed, Z has the further advantage of having its standard error not dependent on the universe value, Z_u, but only on the number of observations on which r is based.[9] The standard error of Z is given by the formula,

$$\sigma_Z = \frac{1}{\sqrt{N-3}} \tag{16}$$

Thus, the Z transformation relieves us from both difficulties we noted—that we needed to know ρ to get σ_r, and that the sampling distribution of r is not symmetrical.

The actual setting up of confidence limits to the estimate of a correlation coefficient in the universe is done in four steps.

1. Find Z corresponding to the absolute value of the observed r from formula (15) or from Appendix Table G.
2. Find σ_Z from formula (16).
3. Find Z_1 and Z_2 by multiplying σ_Z by the number of standard deviation units appropriate for the desired level of confidence (1.96 for 95 percent, 2.58 for 99 percent, etc.) and then adding and subtracting the product of Z.
4. Find r_1 and r_2 corresponding to Z_1 and Z_2 by solving formula (15) for r, or by using Appendix Table G in the reverse direction.

In our example where $r = -.515$ and $N = 31$, these steps for determining the 95-percent confidence limits for r are as follows:

1. $Z = .569511$ (From Appendix Table G)
2. $\sigma_Z = \dfrac{1}{\sqrt{N-3}} = \dfrac{1}{\sqrt{31-3}} = .18896$
3. $Z_{1 \text{ and } 2} = .569511 \pm (1.96)(.18896)$
   $Z_1 = .19915, \quad Z_2 = .93987$
4. $r_1 = -.1966, \quad r_2 = -.7351$

Or to compute the 99-percent confidence limits, we change the multiple in step 3, and repeat steps 3 and 4.

3. $Z_{1 \text{ and } 2} = .569511 \pm (2.58)(.18896)$
   $Z_1 = .08209, \quad Z_2 = 1.05693$
4. $r_1 = -.0819, \quad r_2 = -.7840$

[9] The value of σ_Z is not entirely independent of the universe parameter, Z_u. For a critical appraisal of the Z transformation see Charles C. Peters and Walter R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases* (New York: McGraw-Hill, 1940), pp. 155–157. A different transformation for r is suggested by K. C. S. Pillai, *Sankhya*, 7 (July 1946), pp. 415–422.

It may be noted that we disregarded the sign of r in finding the corresponding value of Z in step 1, and proceeded to find confidence limits as if the observed r were positive until in the final step we attached the original sign of r to r_1 and r_2. This is customary in dealing with the logarithms of negative numbers. Theoretically one cannot express the logarithm of a negative number; in practice, however, one takes the log of the positive number with the same absolute value, performs the necessary operations on it, and after transforming the results back to an antilogarithm, attaches the proper sign.

Either the 95-percent or the 99-percent pair of confidence limits of the correlation coefficient is to be interpreted just as in the case of confidence limits of the mean. We see that even when we choose the less conservative set, the 95-percent confidence limits, we are forced to the conclusion that we do not have a very precise estimate of the degree of association in the universe. The estimate of the coefficient of correlation in the universe is $-.515$ with 95-percent confidence limits of $-.1966$ and $-.7351$, a confidence range of .5385. Let us note what indication there is of asymmetry. The distance between the upper (using "upper" to refer to the limit of greatest absolute value) confidence limit and the estimate of the parameter is .220 while the distance between the lower confidence limit and the estimate of the parameter is .318. Thus, we see that the upper confidence limit is nearer to the estimate than is the lower confidence limit. This shows that any method of setting up confidence limits which involves taking limits equidistant from the estimate would be appreciably inaccurate.

The reliability of coefficients of correlation is a function of the size of the coefficient in the universe sampled and of the size of the sample. Unless the universe coefficient is quite high, however, the number of observations is more important in determining the reliability. Therefore, any estimate of the universe coefficient of correlation based upon observations of the incidence of two characteristics in only 31 areas will have a wide confidence range unless it is a very high coefficient. Because of this, when a coefficient has been determined for a relatively small number of cases—50 or fewer—confidence limits often are not set up; the coefficient is simply tested to discover if it is significantly different from zero.

The total variation in percent of the farms reporting running water in the dwelling in our example is

$$Total\ variation = \Sigma y^2 = 929.828$$

Now of this amount, the proportion,

$$r^2 = (-.515)^2 = .2652 = 26.52\ percent$$

is associated with the variation in the number of nonwhite farm operators per 100 white farm operators. We can actually divide up the variation into its two parts by computing 26.52 percent of the total variation and then by subtracting this amount from the total.

$$\text{Explained variation} = r^2 \Sigma y^2 = (.2652)(929.828) = 246.59$$
$$\text{Unexplained variation} = \text{Total variation} - \text{Explained variation}$$
$$= \Sigma y^2 - r^2 \Sigma y^2 = (1 - r^2) \Sigma y^2$$
$$= (1 - .2652)(929.828)$$
$$= 683.22$$

This is one way of obtaining a meaningful interpretation to the question of how "important" a coefficient with an absolute value of .515 is. About 27 percent of the variation in the percent of farms reporting running water in the dwelling is accounted for or "explained" (in this technical usage) by variation in the number of nonwhite farm operators per 100 farm operators while about 73 percent is unaccounted for. If one prefers to think in terms of the quantity measuring total variation, about 250 of the nearly 1,000 total variation in the percent of farms reporting running water in the dwelling is associated with the ratio of nonwhite to white farm operators.

The above percentages and amounts are definite and precise for the particular set of observations, but they are only estimates of the corresponding universe values. It is not customary to compute the confidence limits of r^2. However, an analysis of variance can be made of the total variation of the Y variable (or of the X variable, if desired) which is an alternate test of the hypothesis that ρ or $\rho^2 = 0$. The above computations can be arranged into the regular form of an analysis of variance. Table 57

Table 57. ANALYSIS OF VARIANCE OF N VARYING UNITS IN CHARACTERISTIC Y FOR TESTING SIGNIFICANCE OF REGRESSION ON CHARACTERISTIC X

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	Ratio of variances, F
Total (Units about $\bar{Y}$).........	Σy^2	$N - 1$		
Regression line on X about $\bar{Y}$............	$r^2 \Sigma y^2$	1	$\dfrac{r^2 \Sigma y^2}{1}$	$\dfrac{r^2(N-2)}{1-r^2}$
Units about regression line................	$(1 - r^2)\Sigma y^2$	$N - 2$	$\dfrac{(1 - r^2)\Sigma y^2}{N - 2}$	

shows the symbols for quantities so arranged and Table 58 shows the actual numbers substituted for the symbols.

The total variation of the Y variable, Σy^2, can be divided into two parts as shown in Table 57; from these parts mean square variances can be computed, and from the mean square variances the ratio of variances or F can be obtained. Inspection of Table 57, however, will show a quicker way of obtaining F if there is no need for the other quantities. In the formula

Table 58. ANALYSIS OF VARIANCE OF 31 ECONOMIC AREAS IN PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING (Y) FOR TESTING SIGNIFICANCE OF REGRESSION ON NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X), 1945

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total (units about $\bar{Y}$)	929.83	30		
Regression line on X about $\bar{Y}$	246.61	1	246.61	
				10.47
Units about regression line	683.22	29	23.56	

$P[F_{1.29} = 10.47] < .01$

Source: Table 55.

for F, Σy^2 appears in both numerator and denominator and can be cancelled out, leaving the relation,

$$F = \frac{r^2(N-2)}{1-r^2} \tag{17}$$

Instead of going through all the computations of Table 58 we can obtain F from substituting values of r and N directly in (17), thus,

$$F = \frac{(-.515)^2(31-2)}{1-(-.515)^2} = 10.47$$

By referring to Appendix Table F, we find that the probability of such an F with one and 29 degrees of freedom is less than .01. This means that in a universe where there is no correlation, the probability that an F as unusual as this would be observed is less than .01. This F test of the significance of the correlation coefficient can be set up in five steps in the same way as any other test of a hypothesis. The only difference is that in step 2, the description of the sampling distribution, we do not actually describe the sampling distribution of F in terms of its mean, standard deviation, and form, but we use Appendix Table F as the description of its distribution.

We shall explain here one other method of testing the hypothesis that

$\rho = 0$ because the test is closely related to the F test. It should be understood that in any practical case it is necessary to use only one of these tests, but since the student may find all of them appearing in the literature, it is a good idea to be familiar with all of them. Furthermore, depending on what further investigations are to be made, sometimes one test is more convenient and sometimes another.

The test to be given now is based on the t distribution, the same distribution used for testing the hypotheses about means. For the correlation coefficient, the hypothesis $\rho = 0$ is tested by computing t from the relation

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{18}$$

Substituting the observed values for r and N in (18), we have

$$t = \frac{(-.515)\sqrt{31-2}}{\sqrt{1-(-.515)^2}} = 3.235$$

From Appendix Table D we find that the probability of a $t = 3.235$ with 29 degrees of freedom is less than .01.

For convenience of reference, as well as for comparison of the results for this particular example, the following summary of the three tests of significance of the correlation coefficient is given, showing the argument necessary to enter the table, the degrees of freedom, the table used, and, for our example, the actual probability determined by the test.

SUMMARY OF THREE TESTS OF THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT (TESTS OF THE HYPOTHESIS THAT $\rho = 0$) FOR PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING AND NUMBER OF NON-WHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS, 1945, FOR 31 ECONOMIC AREAS

Argument needed to enter table	Degrees of freedom	Table describing distribution	Probability that a value so unusual would occur if $\rho = 0$
$\dfrac{r}{\sigma_r} = r\sqrt{N-1}$		Areas under the normal curve (Appendix Table A)	$P = .0048$
$F = \dfrac{r^2(N-2)}{1-r^2}$	$n_1 = 1$ $n_2 = N-2$	Distribution of F (Appendix Table F)	$P < .01$
$t = \dfrac{r\sqrt{N-2}}{\sqrt{1-r^2}}$	$N-2$	Distribution of t (Appendix Table D)	$P < .01$

It can be seen in the above summary that the expression for t is the square root of the expression for F. If the tables of the distributions of F and t are examined, it will be found that the value of t for n degrees of freedom is equal to the square root of the value of F when $n_1 = 1$ and $n_2 = n$. Therefore, the F and t tests summarized above are really identical, and as would be expected, they yield the same probability in the last column. Although it cannot be seen in this case, the first test yields a slightly higher probability than the other two tests. We say that the first test is more "stringent" than the others. The difference in results is due to the incorrectness of certain assumptions on which the first test is based. If the difference is appreciable, the results of the F or t test are to be accepted instead of the results of the first test.

The standard error of estimate.   We shall develop one other measure based upon the analysis of variance of the Y variable, which may also contribute in interpreting the meaning of the *degree* of association. We referred earlier to a measure of the scatter of the dots around the regression line in Figure 39. Analogous to a standard deviation of measures from their mean, there is a standard error of estimate of measures from a regression line. The formula for the standard deviation, it will be remembered, can be written

$$s = \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{N}}$$

The similarity of the measures is shown by the similarity of the above formula to that for the standard error of estimate, which is

$$s_{y_s} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N}} \tag{19}$$

Formula (19) affords a descriptive measure of the standard error of estimate of these observed 31 dots around the line. It is almost never used, however, for we are primarily interested in the standard error of estimate for the universe, since only for cases other than the ones observed would one have occasion to use the regression equation for estimating or predicting values of Y from a knowledge of corresponding values of X. Since two degrees of freedom have been sacrificed in determining the regression line, the denominator in (19) should be $N - 2$ for the estimate of the universe value of the standard error of estimate about the regression line, thus,

$$\hat{\sigma}_{y_s} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N - 2}} \tag{20}$$

It has been pointed out that the numerator of the fraction under the radical is called the "unexplained" variation. Since it is a measure of the *scatter* of the observed points about the regression line, we denote the unexplained variation as Σy_s^2 and determine it by evaluating the formula,

$$\Sigma y_s^2 = \Sigma(Y - Y_c)^2 = (1 - r^2)\Sigma y^2 \tag{21}$$

Substituting (21) in (20) we have

$$\hat{\sigma}_{y_s} = \sqrt{\frac{(1 - r^2)\Sigma y^2}{N - 2}} \tag{22}$$

Referring to Table 57, we see that this quantity is simply the square root of the "mean square variance" of units about the regression line. Therefore, from Table 58 we can take directly the value of the expression under the radical and find for our example that the estimate of the standard error of estimate in the universe is

$$\hat{\sigma}_{y_s} = \sqrt{23.56} = 4.854$$

For comparison with this standard error of estimate, we may also compute from data in Table 58 the estimate of the universe value of the standard deviation of the Y variable when considered as a single distribution from formula (3) of Chapter 16.

$$\hat{\sigma}_y = \sqrt{\frac{\Sigma y^2}{N - 1}} = \sqrt{\frac{929.83}{30}} = 5.567$$

The value of $\hat{\sigma}_{y_s} = 4.845$ describes the scatter of the observations around the regression line analogously to the way the value of $\hat{\sigma}_y = 5.567$ describes the scatter of the observations around the mean of the Y distribution. This may be interpreted to mean that if the association of characteristic Y with characteristic X is of a degree given by the measure $r = -.515$, actual values of Y would deviate from estimates of Y made from a knowledge of corresponding X values and a knowledge of the regression equation of Y on X with a scatter measured by the standard error of estimate of 4.854; whereas, the actual values would deviate from a constant estimate of Y, $\bar{Y}$, with a scatter measured by the standard deviation of 5.567. Foreknowledge of X improves the estimates of Y, but not very greatly.

Nature of association in the universe. The nature of association in the sample has been described by the regression equation (10),

$$Y_c = 17.95 - .05338X \tag{10}$$

We have remarked that the important coefficient of this equation is $b = -.05338$. This regression coefficient specifies the number of units of

change in Y associated with a unit increase in X. The estimate of the value of b in the universe, which we shall designate as b_u is,[10]

$$\hat{b}_u = b = -.05338 \tag{23}$$

If the correlation between two characteristics is zero, b is zero, because no different levels of Y are associated with different levels of X. In fact since

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \qquad \text{and} \qquad b = \frac{\Sigma xy}{\Sigma x^2}$$

we can write the formula for b in terms of r, thus,

$$b = r\sqrt{\frac{\Sigma y^2}{\Sigma x^2}} = r\frac{s_y}{s_x} \tag{24}$$

and from this formula we can see that when $r = 0$, b must also be equal to zero, and that if $r \neq 0$, then b must also be unequal to zero, unless the X variable shows no variation at all (in which case the methods of correlation analysis are not appropriate).

Therefore, any of the tests of significance of the correlation coefficient are at the same time tests of the significance of the regression coefficient. By making these tests we have already rejected the hypothesis that $b_u = 0$. The sampling distribution of b_u like that of $\hat{\rho}$ is dependent upon the unknown universe values of ρ and is, therefore, difficult to deal with.[11]

Again we shall summarize the information we have obtained from investigating the four aspects of association in the infinite universe of possibilities, from which the 31 observations may be considered a random sample.

SUMMARY OF DESCRIPTION OF ASSOCIATION BETWEEN PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING (Y) AND NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X), 1945, FOR THE INFINITE UNIVERSE OF POSSIBILITIES FROM WHICH THE 31 OBSERVATIONS ON THE ECONOMIC AREAS MAY BE CONSIDERED A RANDOM SAMPLE

Aspect of association	Description
1. *Existence of association*	We have established the fact of the existence of association in the universe by finding grounds to reject the hypothesis that $\rho = 0$. (This is usually reported by the simple statement that "r is significant.")

[10] We do not use β for the regression coefficient in the universe because β_1 and β_2 are used as symbols to define certain functions of moments (see pages 211–212) while the letter without a subscript is used for the regression coefficient when both variables are measured in terms of their standard deviation units.

[11] See Tippett, *op. cit.*, pp. 181–188; and R. A. Fisher, *Statistical Methods for Research Workers*, 10th ed. (London: Oliver and Boyd, 1946), pp. 131–140.

Aspect of association	Description
2. *Direction of association*	The test of the hypothesis described in 1 also indicates that ρ has the same sign as r; therefore, we have established the fact that the association in the universe is negative.
3. *Degree of association*	Our estimate of the degree of association in the universe is given by our estimate of $\hat{\rho} = -.515$, with 95-percent confidence limits of $-.197$ and $-.735$. The estimate of the universe value of the standard error of estimate of the regression equation is a function of the degree of association and is $\hat{\sigma}_{y_e} = 4.854$, which may be compared with $\hat{\sigma}_y = 5.567$.
4. *Nature of association*	(Under the assumption of linearity of form) our estimate of the nature of the association in the universe is given by the regression equation, $$Y_c = 17.95 - .05338\,X$$ While we did not attempt to set up confidence limits for $\hat{b}_u = -.05338$, the tests in 1 also establish the fact that b_u is different from zero and negative.

COMPUTATIONS OF CORRELATION AND REGRESSION COEFFICIENTS FROM GROUPED DATA

If the number of paired observations which constitute the data for a correlation problem is greater than 200 or 300, it will be found to be time saving to compute r's and b's and a's from grouped rather than ungrouped data. We shall explain the computations for these coefficients using the same data for which the methods for ungrouped data have been illustrated in order that the student may see the two methods of computation applied to the same problem. Since the interpretations of the coefficients are exactly the same, we shall not have to repeat them, and can focus on the mechanics of computation. This focusing on computations is desirable, for at first glance a correlation computation table such as Table 59 looks very complex, and it is well not to attempt to master simultaneously the mechanics of computation from grouped data and the meaning of the coefficients for the first time.

Required data. If data are available in the form of Table 55, the first step is to select class intervals for each of the characteristics and then to cross-tabulate the data into a table of the form of Table 56. A table of data so cross-tabulated is often called a "correlation table," although some reserve this term for a table such as Table 59, which contains all the

Table 59. COMPUTATIONS FOR OBTAINING CORRELATION AND REGRESSION COEFFICIENTS FROM GROUPED DATA

Class limits	Number of nonwhite farm operators per 100 white farm operators				$X \rightarrow$	0–24	25–49	50–74	75–99	100–124	125–149	150–174	175–199	Sums
Percent of farms reporting running water in dwellings Y ↓	$m \rightarrow$					12.0	37.0	62.0	87.0	112.0	137.0	162.0	187.0	
	↓	$d' \rightarrow$				−2	−1	0	1	2	3	4	5	
		↓	$f \rightarrow$			8	8	5	3	3	1	0	3	31
			↓	$fd' \rightarrow$		−16	−8	0	3	6	3	0	15	3
				↓	$f(d')^2$	32	8	0	3	12	9	0	75	139
27.5–29.9	28.7	6	1	6	36	1 −12 −12								$\Sigma fd'x\,d'y$ −12
25.0–27.4	26.2	5	1	5	25	1 −10 −10								−10
22.5–24.9	23.7	4	0	0	0	0								
20.0–22.4	21.2	3	3	9	27	1 −6 −6	1 −3 −3		1 3 3					−6
17.5–19.9	18.7	2	5	10	20	2 −4 −8	3 −2 −6							−14
15.0–17.4	16.2	1	3	3	3	1 −2 −2		2 0 0						−2
12.5–14.9	13.7	0	8	0	0	2 0 0	1 0 0	1 0 0	1 0 0	2 0 0			1 0 0	0
10.0–12.4	11.2	−1	1	−1	1								1 −5 −5	−5
7.5–9.9	8.7	−2	7	−14	28			3 2 6	2 0 0		1 −4 −4		1 −10 −10	−8
5.0–7.4	6.2	−3	2	−6	18					1 −3 −3		1 −9 −9		−12
Sums			31	12	158									−69

Source: Table 56.

information of Table 56 and in addition certain computations basic to obtaining the correlation and regression coefficients.

Let us emphasize once more that data given as single distributions

are *not* sufficient for correlation analysis. To investigate association we must know not only the number of observations in each class interval of one variable, but also we must know the class interval of the other variable into which each of these observations falls. The practical reason for which this point has been reiterated is as follows: If associations are to be investigated between two characteristics in a study involving firsthand collection of data, they should be planned for before the assembling and tabulating stages in order to secure the necessary cross-tabulations. If a study uses a long schedule and machine tabulation, it may be necessary to punch the data from one schedule onto more than one card. In such a case items can be cross-tabulated *only* if they are on the same card. If the later analysis is not planned in advance of punching the cards, one may find it impossible to investigate the association between two characteristics on different cards, although he will be able to get frequency tables for each characteristic separately.

Preparation of correlation computation table. Table 59 can be recognized as an extension of Table 56. In fact, the upper entry in each cell of the body of Table 59 is the frequency appearing in the corresponding cell of Table 56. The top six rows of Table 59 and the leftmost six columns are not in the body proper of the table. The six rows contain certain information for the X distribution while the six columns contain the same type of information for the Y distribution, and the nature of the entries in both rows and columns are specified by the words or symbols in the first six cells along the diagonal beginning at the upper left corner of the table. These designations are in order as follows: class limits; midvalues of class intervals, m; step deviations, d'; frequencies, f; products of frequencies and step deviations, fd'; and products of frequencies and squared step deviations, $f(d')^2$. These are the same quantities which would be computed for determining the means and the standard deviations of the single distributions by the short method with grouped data. Filling in these six rows and columns involves nothing new, nothing which has not been explained in Chapter 9—it is only the arrangement of work that is different.

The new computations of Table 59 are those designed to provide data on the *covariation* of the two characteristics. These computations are made as follows:

1. In each cell of the body of the table where a frequency is entered, write just underneath the frequency (preferably in pencil with a colored lead) the number which is the product of the d' of the Y variable and the d' of the X variable for that cell. For example, in the first row and first column of the body of Table 59, the number -12 is secured by multiplying $(-2)(6) = -12$. These products will be designated as $d'_X d'_Y$.

2. In each cell of the body of the table where a frequency is entered, write

just underneath the $d'_xd'_Y$ entry the number which is the product of the other two entries of the cell. (Still another color of lead is advised for this.) For example, in the fifth row and first column of the body of Table 59, the number -8 is secured by multiplying $(2)(-4) = -8$. These products will be designated as $fd'_xd'_Y$.

3. Sum the $fd'_xd'_Y$'s for each row, entering their sum in the last column of the table. Add these sums and enter the grand total in the extreme lower right cell of the table, labeling this sum as $\Sigma fd'_xd'_Y$.

The correlation computation table is now complete. From the sums of such a table one can compute r, b, a, and s_{ys} to describe the association of the two characteristics and, of course, $\bar{X}$ and s for each characteristic separately. From Table 59 the data which are to be used for computing the correlation and regression coefficients are the following six sums:

$$\Sigma f_X = \Sigma f_Y = N = 31 \qquad\qquad \Sigma fd'_xd'_Y = -69$$
$$\Sigma f_xd'_x = 3 \qquad\qquad\qquad\quad \Sigma f_Yd'_Y = 12$$
$$\Sigma f_x(d'_x)^2 = 139 \qquad\qquad\quad \Sigma f_Y(d'_Y)^2 = 158$$

In addition, one must know the location of the guessed means and the size of the class intervals, which in Table 59 are the following:

$$\bar{X}' = 62.0 \qquad\qquad\qquad \bar{Y}' = 13.7$$
$$i_X = 25.00 \qquad\qquad\qquad i_Y = 2.50$$

Computation of r. The basic formula for the coefficient correlation given earlier in this chapter is

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \tag{4}$$

The following alternate form was also given as equivalent to (4),

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \tag{5}$$

In formula (5) the numerator is equal to N times the covariation and the denominator is equal to the geometric mean of N times the two total variations. For use with grouped data formula (5) is modified to the following form,

$$r = \frac{N\Sigma fd'_xd'_Y - (\Sigma f_xd'_x)(\Sigma f_Yd'_Y)}{\sqrt{[N\Sigma f_x(d'_x)^2 - (\Sigma f_xd'_x)^2][N\Sigma f_Y(d'_Y)^2 - (\Sigma fd'_Y)^2]}} \tag{25}$$

The equivalence of this formula to formula (5) is suggested by the notation. We now substitute the six sums listed above in this formula and thus compute r from grouped data.

$$r = \frac{(31)(-69) - (3)(12)}{\sqrt{[(31)(139) - (3)^2][(31)(158) - (12)^2]}} = -.493$$

The value obtained for the correlation coefficient is slightly smaller than that obtained from ungrouped data because the grouping of data tends to exaggerate the amount of variation in each single distribution, thus making the denominator slightly too large. Sheppard's corrections may be applied in the denominator of (25) to prevent this exaggeration. Unless the number of cases is very great, however, the use of such a correction is not advised.

Computation of b. The basic formula for the regression coefficient given earlier in this chapter is

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} \tag{8}$$

The following alternate form was also given as equivalent to formula (8),

$$b_{YX} = \frac{N\Sigma X Y - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \tag{8A}$$

In formula (8A) the numerator is equal to N times the covariation and the denominator is equal to N times the total variation of X. For use with grouped data formula (8A) is modified to the following form,

$$b_{YX} = \frac{N\Sigma f d'_X d'_Y - (f_X d_X)(f_Y d_Y)}{N\Sigma f_X (d'_X)^2 - (\Sigma f_X d'_X)^2} \times \frac{i_Y}{i_X} \tag{26}$$

The equivalence of formula (26) to formula (8A) is suggested by the notation. We now substitute the sums and class interval sizes from p. 439 and, thus, compute b from grouped data,

$$b_{YX} = \frac{(31)(-69) - (3)(12)}{(31)(139) - (3)^2} \cdot \frac{2.50}{25.00} = -.0506$$

The formula for a given previously is

$$a_{YX} = \frac{\Sigma X - b_{YX}\Sigma X}{N} \tag{9}$$

This can be written in an equivalent form, thus,

$$a_{YX} = \bar{Y} - b_{YX}\bar{X} \tag{9A}$$

For computing $\bar{X}$ and $\bar{Y}$ to substitute in this formula we use the regular formulas given in Chapter 8 for computing means by the short method with grouped data,

$$\bar{X} = \bar{X}' + \frac{\Sigma f d'}{N} i \tag{27}$$

If we substitute the right member of (27) for $\bar{X}$ and $\bar{Y}$ in formula (9A), we obtain

$$a_{YX} = \bar{Y}' + \frac{\Sigma f_Y d'_Y}{N} i_Y - b_{YX}\left(\bar{X}' + \frac{\Sigma f_X d'_X}{N} i_X\right) \qquad (28)$$

We now substitute the data from page 439 and the value of b in (28) to compute a_{YX}, thus,

$$a_{YX} = 13.70 + \frac{12}{31}(2.5) - (-.0506)\left[62.0 + \frac{3}{31}(25)\right] = 17.93$$

Writing the values of a and b just found as the coefficients in a regression equation, we have

$$Y_c = 17.93 - .0506\,X \qquad (29)$$

which may be compared with the similar equation (10) secured from ungrouped data.

Computation of other measures relating to association from grouped data. Any of the three tests of significance, in addition to the Z transformation for setting up confidence limits can be made with only N and r as data. Since we have shown how to compute r for grouped data, we do not need to go through the procedures again.

If one wishes to make an analysis of variance, such as that shown in Table 58, however, or to compute the standard error of estimate of the regression equation, he must have in addition to the above measures, that is, in addition to r and N, a measure of the total variation of the Y variable. This may be secured from the sums and other data of page 439. We have pointed out that the expression

$$N\Sigma f_Y(d'_Y)^2 - (\Sigma f_Y d'_Y)^2$$

is equal to N times the total variation of the Y variable expressed in class intervals. Therefore, if we divide this expression by N and multiply it by the square of the class interval, we will have the desired quantity. This may be expressed by formula thus,

$$\Sigma y^2 = [N\Sigma f_Y(d'_Y)^2 - (\Sigma f_Y d'_Y)^2]\frac{i_Y^2}{N} \qquad (30)$$

Substituting the data of page 439 into (30), we have

$$\Sigma y^2 = 31(158) - (12)^2\frac{.(2.5)^2}{31}$$

$$= 958.47$$

This value for Σy^2 is slightly larger than 929.83, the corresponding measure of the total variation secured from ungrouped data, as would be expected since grouping exaggerates variation. Again we mention that if the number

of cases is very large, one might apply Sheppard's correction to prevent this exaggeration. From the measure of total variation just secured one may proceed with analysis of variance or may compute the standard error of estimate exactly as before. A formula analogous to (30) can be secured for Σx^2 by simply substituting X for Y in the subscripts, if one wishes to analyze the variance of the X characteristic. If for any reason one wishes to compute the quantity Σxy from grouped data, the following similar formula can be used,

$$\Sigma xy = [\, N\Sigma fd'_X d'_Y - (\Sigma f_X d'_X)(\Sigma f_Y d'_Y)\,]\frac{i_X i_Y}{N} \qquad (31)$$

FORM OF ASSOCIATION

Meaning of linear form. Thus far in the treatment of correlation it has been assumed that the relationship being investigated between the two characteristics is of linear form. The coefficient of correlation, r, measures the closeness with which the dots of Figure 39 cluster about a *straight line;* the regression equation determined by the coefficients a and b describes the *straight line* which best fits the dots of Figure 39. In linear correlation analysis we assume that b units of change in the Y characteristic is associated with one unit change in the X characteristic over the entire observed range of values of X; we assume that if there is a "law" of relationship between the two characteristics, it can be formulated as an equation of the first degree.

Meaning of nonlinear form. In the realms of study of physical characteristics where measurement is highly developed it has been found that characteristics are often associated in such a way that a first degree equation does not satisfactorily formulate their "law" of relationship because the amount of change in the Y characteristic corresponding to a unit change in the X characteristic does not remain the same throughout the range of observed X values. The volume of a cube, for instance, varies as the *third* power of its edge; the rate of heat transmission varies as the *fourth* power of the difference in temperature of the two bodies; the yield of wheat increases with increases in fertilizer up to a certain stage and then decreases with increase in fertilizer; the weight of a human being increases very rapidly for a certain period of time, then increases less rapidly, and finally reaches a standstill. In all of these illustrations the form of association between the two characteristics is nonlinear.

Regression equations as scientific laws. In the realm of characteristics of interest to sociologists it is probable that so many factors are involved in determining the incidence of any one characteristic that we shall never arrive at formulations of "laws" of relationship between any

two characteristics which will be fitted by observations so well as are the laws of the physical sciences. Therefore, we do not usually call linear or nonlinear regression equations "laws," since the concept of "scientific law" means to most people a more precise description of relationship than can be made for sociological characteristics. And yet, it must be recognized that correlation analysis, or more acurately, regression analysis, is an attempt to formulate from observed data a rough approximation of a law of relationship between characteristics, since after all a scientific "law" is simply a description of relationships. In assuming a linear form of regression one assumes that the first power of one characteristic is (on the average) always equal to some constant plus a multiple of the first power of the other characteristic in the approximation to the law.

Nonlinear regressions in sociological research. There are many relationships between sociological characteristics for which this assumption as to form leads to absurd results. Many are of the nature of the yield-fertilizer relationship, where we say that after a certain stage the "law of diminishing returns" sets in. Many others are of the nature of the weight-time relationship, where an upper asymptote is approached. Still others are of the sort where projection of the straight line beyond the observed range leads to negative or infinite results which may be absurd or impossible.

Therefore, sociologists do not always restrict themselves to linear equations in formulating their approximations to laws of relationship from observed data. Especially where there is some theory as to the relationship investigated, from which one can deduce some form of relationship other than linear, does the sociologist abandon the assumption of linearity. In such a case a nonlinear equation is formulated by methods known as curve fitting. (Since mathematically a straight line is one type of curve, the fitting of a regression line is also a case of curve fitting.) Curves (in the everyday sense of the word) can be fitted to describe the nature of the relationship between two characteristics by procedures similar to those used in fitting a straight line, but more elaborate. After the curve is fitted, coefficients analogous to r, r^2, and $\acute{\sigma}_{y_s}$ can be computed to measure the closeness with which the observations cluster around the curve.

Reasons for the assumption of linearity in practical work. Linear correlation and regression, however, are much more commonly used than nonlinear in sociological research, even when the assumption of linearity is not logically justifiable. There are a number of reasons for making the assumption of linearity in correlation and regression analysis.

One reason is that linear coefficients are more easily computed; the arithmetic of linear methods is simpler than that of nonlinear.

Another reason is that the linear coefficient of correlation is always

smaller than any curvilinear coefficient (if a curvilinear assumption is justified), and if the emphasis of the investigation is on the first aspect of association, that is, existence, a test of significance of r which establishes the fact of the existence of association in the universe may be all that is desired. (However, if the test shows that r is not significantly different from zero, this does *not* mean that there may not be curvilinear association in the universe.)

Another reason is that even though there may be interest in describing all aspects of relationship, the interest may be restricted to an empirical description of the aspects within the observed range of the measures of the characteristics, within which the relationship may be nearly linear, even though it is curvilinear beyond this range. Segments of a curve, if they are short, often approach straight lines very closely, and, therefore, straight lines can be used as first approximations to segments of curves.

Another reason is that often there is no previous theory which can enable one to deduce in advance the form of relationship to be expected. In such a case the simplest methods are always tried first.

Still another reason is that even though nonlinearity may be anticipated or suspected, the proof for nonlinearity requires the computation of linear coefficients to be used in testing the significance of departure of nonlinear coefficients from them.

A final reason is that a regression equation involving only two constants, such as a linear equation, is a more "efficient" statistical description than one involving more than two, as many nonlinear equations do. One of the primary purposes of scientific research is to reduce numerous observations to some sort of order which can be expressed by fewer numbers. The fewer numbers used in description, the more efficient is the statistical device.

In the light of these reasons for using linear correlation and regression, we offer the following suggestions as to when to assume linearity in practical procedures. If the construction of a scatter plot is made the first step in correlation analysis, any marked departure from linearity will be quickly apparent as soon as one becomes experienced in interpreting a scatter plot. Unless nonlinearity is evident from a scatter plot or unless there is interest in fitting some theoretically deduced nonlinear form, the assumption of linearity as to form of the relationship *within the observed range* is the advised procedure. Extrapolation of the description of the existence, direction, degree, or nature [12] of the association beyond the

---

[12] The description of the "nature" of an association includes the specification of both the form and the constants of the equation relating the two characteristics. Heretofore, we have assumed the form to be linear and have confined our investigation of "nature" of association to determining the constants for the linear equation. In the next section we shall learn how to test such an assumption of linearity. When such a test is made, it is reported under the aspect of "nature" in a summary of the description of association.

range of the observations must be justified for each particular case. Extrapolation of a purely empirical description of association is, of course, not so easy to justify as is extrapolation of a description based upon a theoretically deduced form of relationship.

NONLINEAR CORRELATION AND REGRESSION

In correlation and regression analysis nonlinear methods are indicated when the investigation is made for the purpose of checking the validity of some theoretically deduced nonlinear form of association with observations or when the scatter plot suggests a nonlinear form of association for purely empirical description of the relationship. We shall discuss briefly the methods employed in each of these cases and the principles upon which they are based, and then we refer the reader to advanced treatments on curve fitting and nonlinear analysis for the actual computational procedures.

Fitting of theoretically deduced nonlinear forms to observed data. One of the commonest types of problems with a theoretically deduced form of association is in growth problems. Growth curves are the regression equations between two characteristics, one of which (Y) is the cumulative size or number or amount of something, the other of which (X) is time, measured in hours, days, years, decades, or any appropriate time units. Not only growth curves but also every time series can be thought of as a correlation problem since it displays simultaneously the measures on two characteristics. The closeness with which the observations of a time series cluster about any equation fitted to them can be measured by a coefficient of correlation. This is not usually done in analysis of economic time series, however, for the principal interest of the investigation is more likely to be centered on the cyclical or seasonal fluctuations of the observations around the line or curve fitted rather than on the line or curve itself. In analysis of the type of time series known as a growth curve, however, primary interest is often in finding how closely observations correspond to some theoretically deduced form of a "law of growth." In such a case the coefficient of curvilinear correlation and the related standard error of estimate may be used in their descriptive function as measures of the closeness of fit of observations to a curve, and even to compare closeness of fit for two different forms of fitted curves. The measures are not to be used in their generalizing function without qualifications, however, because of the lack of independence of the successive observations, which exaggerates the degrees of freedom and hence underestimates measures of error.[13]

Other than growth curves, few theoretically deduced forms have been

[13] See page 352.

fitted to sociological data. An exception is the fitting of a theoretically deduced form to migration data by Samuel A. Stouffer.[14] As sociological theory develops more precise formulations, we can expect more examples of determining from observed data the constants for a theoretically deduced form of curve in many other fields than growth curves. As theory develops even further, it may be that not only the form of the association but also the parameters (universe values for regression coefficients) may be deduced for many relationships, and the problem will then become the testing of the closeness of fit of observations to theoretically derived equations.

Fitting of empirically determined nonlinear forms to observed data. When there is no theory to suggest the form of the equation which should be fitted, one simply looks at the scatter plot, guesses which one of several relatively simple forms would best describe the observed association, and tries it out. To be able to make a good guess, he should know what geometric form corresponds to the commonly used algebraic forms. Figure 40 shows a number of relatively simple types of curves with different constants. The first is a reciprocal curve of the form,

$$Y_c = \frac{1}{a + bX} \tag{32}$$

The second is a modified exponential curve of the form,

$$Y_c = a + b^X \tag{33}$$

The third is a logarithmic curve of the form,

$$Y_c = a + b \log X \tag{34}$$

The fourth is of a second degree polynomial of the form,

$$Y_c = a + bX + cX^2 \tag{35}$$

The fifth is of a third degree polynomial of the form,

$$Y_c = a + bX + cX^2 + dX^3 \tag{36}$$

These five series are only suggestive; there is no limit to the variety of forms which may be fitted. In any of the above forms X and Y may be interchanged.

The problem of nonlinear regression involves as before the finding of the values of the two, three, or more coefficients of the equation of the form chosen. Again the determination of the values of the constants

---

[14] Samuel A. Stouffer, "Intervening Opportunities: A theory Relating Mobility and Distance," *American Sociological Review*, 5 (December 1940), pp. 845–867.

which will make the chosen form of equation best fit the observed data is analogous to the determination of the values of a and b in linear regression.

Now let us consider how we may define a coefficient of nonlinear correlation analogous to r in linear correlation. It will be remembered that r^2 is the proportion that the "explained" variation is of the total variation of the Y distribution. For further work it is more convenient to use a corollary to this relation, that is, that $1 - r^2$ is the proportion

Figure 40. TYPES OF NONLINEAR FORMS WHICH MAY BE FITTED TO OBSERVED DATA.

the "unexplained" variation is of the total variation. Expressed in symbols for a linear regression where

$$Y_c = a + bX$$

this corollary becomes

$$1 - r^2 = \frac{\Sigma(Y - Y_c)^2}{\Sigma(Y - \bar{Y})^2} = \frac{\Sigma y_s^2}{\Sigma y^2} \tag{37}$$

Now we can define analogously a coefficient of curvilinear correlation. r_c,[15] for any nonlinear regression where[16]

---

[15] Most authors use the symbol ρ for the coefficient of curvilinear correlation, but we have already used ρ to denote the coefficient of correlation in the universe; hence the symbol above.

[16] Where $f(X)$ means any single valued function of X—that is, it designates some equation by which for any given value of X, a corresponding value of Y_c may be determined.

$$Y_c = f(X)$$

by the relation,

$$1 - r_c^2 = \frac{\Sigma(Y - Y_c)^2}{\Sigma(Y - \bar{Y})^2} = \frac{\Sigma y_s^2}{\Sigma y^2} \tag{38}$$

For actual computation of r_c^2, from which r_c is obtained by extracting the square root, the denominator is secured just as for linear coefficients. The numerator can be secured in several ways. The most obvious way, although not the most convenient one, is to compute for each observed Y its corresponding Y_c (the Y_c which has the same X value), subtract the latter from the former, square the difference $(Y - Y_c = y_s)$, and add these squared differences for all observed values of Y. Shorter and more convenient methods differ for each form of the regression equation, but they are analogous to the computation of $\Sigma(Y - Y_c)^2$ in linear regression from the formula,

$$\Sigma(Y - Y_c)^2 = \Sigma Y^2 - (a\Sigma Y - b\Sigma XY) \tag{39}$$

The coefficient of curvilinear correlation, r_c, can be interpreted as the square root of the proportion of variation in Y "explained" or accounted for by the curvilinear regression,

$$Y_c = f(X)$$

Unlike the coefficient of linear correlation, which is symmetrical with respect to the two variables, r_c is the square root of the proportion of variation in Y explained by its association with X, but it is *not* (except in limiting cases) the proportion of variation in X explained by the curvilinear regression,

$$Y_c = f(X) \quad \text{or} \quad X_c = f(Y)$$

The coefficient r_c should perhaps be written $r_{c_{yx}}$ to denote that it measures the closeness of the curvilinear regression of Y on X, but as with the coefficient b in linear regression, we are omitting the subscripts with the understanding that we mean the regression of Y on X unless otherwise specified.

If for a given example r_c is larger than r, we say that the curvilinear equation describes the observed association better than the linear equation does. If, however, the curvilinear equation has more than two constants which have to be determined from the observed data, the loss in degrees of freedom may not be compensated for by the increase in size of the correlation coefficient. For instance, if a curve whose form requires the evaluation of N constants is fitted to a series of N paired observations on X and Y, it can be made to go through every single one of the obser-

vations, that is, to fit the observations perfectly with an r_c equal to one. This is true even though the universe from which the N observations may be considered a random sample has an r_c equal to zero. Thus, it is evident that degrees of freedom must be taken into consideration in testing whether or not r_c is significantly higher than r. In such tests it is r_c^2 and r^2 which are usually compared, and this may be done by the use of analysis of variance, which takes into account the number of degrees of freedom.

The correlation ratio. If there is no theory to suggest a specific nonlinear form, the extent of our interest in curvilinearity may be simply to determine whether the nature of the association between two variables departs significantly from linearity. Suppose that there is no interest in the form other than discovering whether it is linear or nonlinear. Then without exploring which form would give the best fit, we can determine what is called the "correlation ratio" and test to see if its square is significantly higher than r^2.

The correlation ratio, E, is a special case of r_c and is again defined analogously to r as the square root of the ratio of the variation in Y "explained" to the total variation of Y. The use of the word "explained" here is not so easy to make clear as before, but let us go on and again define the quantity $1 - E^2$ as the ratio of the "unexplained" to the total variation in Y, where "unexplained" variation means the variation of the Y's in each class interval of X around the mean Y value for that class interval of X. It is immediately apparent that the correlation ratio has meaning only when the data are grouped according to class intervals of X, although they may be grouped or ungrouped with respect to Y. Let us consider the case where they are ungrouped with respect to Y. We may define the correlation ratio E by the relation,

$$1 - E^2 = \frac{\Sigma(Y - \bar{Y}_x)^2}{\Sigma(Y - \bar{Y})^2} = \frac{\Sigma(Y - \bar{Y}_x)^2}{\Sigma y^2} \tag{40}$$

Where $\bar{Y}_x$ is the mean of the Y values in each class interval of X, often called the "column mean" with reference to a correlation table. It is evident that E is equivalent to an r_c computed with respect to a regression equation with as many constants as there are column means, fitted so that the curve goes through each column mean. Such an equation would probably not be interpretable theoretically, but the closeness of its fit to the observed points represents a limit to the closeness of fit which can be expected from *any* nonlinear regression with m constants (or fewer than m constants) where m is the number of column means or the munber of class intervals of the X variable. If we increase the number of class intervals of X and let m approach N, E will approach the value one. It can be proved that E is a function not only of the degree of nonlinear associa-

tion in the universe, but also of the arbitrarily selected number of class intervals, m. Therefore, it should be used as a measure of degree of curvilinear association only when m is small in comparison with N.

Example of computation of the correlation ratio. Let us examine again the scatter plot of the 31 observations on the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators in 1945 shown in Figure 39. When the correlation is no greater in absolute value than .515, it is not always easy to tell from inspection if the regression departs significantly from linearity. Let us next look at Table 60, a hybrid table based on the

Table 60. PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING (Y) FOR 31 ECONOMIC AREAS GROUPED BY CLASS INTERVALS OF NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X), 1945

	Nonwhite farm operators per 100 white farm operators							
	0–24	25–49	50–74	75–99	100–124	125–149	150–174	175–199
	28.0	20.2	17.3	20.6	13.5	5.8	—	14.4
	27.4	19.6	15.2	12.8	13.5			10.6
	20.4	18.7	12.5	7.3	9.0			9.7
	18.6	17.5	9.6					
	18.3	13.9	7.9					
	16.8	9.7						
	14.3	8.5						
	12.9	7.8						
Sums	156.7	115.9	62.5	40.7	36.0	5.8		34.7
k_i	8	8	5	3	3	1		3
$\bar{Y}_x$	19.59	14.49	12.50	13.57	12.00	5.80		11.57

Source: Table 55.

data of Table 55 with X values grouped according to the same class intervals used in Tables 56 and 59 and with Y values ungrouped. Figure 41 corresponds to Table 60 in the same way that Figure 38 corresponds to Table 55. In Figure 41 the X coordinate of each point is taken to be the midvalue of the class interval within which it falls, while the Y coordinate is the observed ungrouped Y value. The mean Y value for each column of Table 60 (each class interval of the X variable) is shown as a circle in Figure 41. The variation of the Y's in each class interval about their column mean is the quantity referred to as the "unexplained" variation in the definition of the correlation ratio.

If the mean of each column is actually computed, as has been done and is shown in Table 60, the numerical value of the unexplained variation

can be found by subtracting from each Y value the mean of its column, squaring the difference, and adding these squares to obtain the expression,

$$\text{Unexplained variation} = \Sigma(Y - \bar{Y}_x)^2$$

which can then be used as the numerator of (40) to compute E.

Such a procedure would be laborious, however, and because it involves the use of rounded means, it would not give very accurate results unless computations were taken to many decimal places. If one studies carefully the problem involved, he can see that it is identical with the

Figure 41. Mean Percent of Farms Reporting Running Water in Dwelling ($\bar{Y}_x$) of Economic Areas Grouped by Class Intervals of Number of Nonwhite Farm Operators per 100 White Farm Operators (X), 1945. (Source: Table 60.)

problem of computing the "within-class" variation of analysis of variance problem. In fact, the square of the correlation ratio is simply the ratio of the "between-class" variation to the "total" variation in an analysis of variance problem where the "classes" are class intervals of a second quantitative characteristic. The correlation ratio and the test of significance of departure from linearity based upon it are an example of the use on a quantitative characteristic of methods primarily designed for a nonquantitative characteristic, the class intervals of the quantitative characteristic being treated as categories of the nonquantitative characteristic.

Therefore, for computation procedures, we refer to the computation guide for analysis of variance with unequal class frequencies given in

Chapter 22, page 402. We note that we must adapt the formulas of the computation guide to the notation of the present problem. The first adaptation is to change the X's to Y's, since in the present problem it is the variation of the Y's which we are investigating. We can find then that the square of the correlation ratio can be computed from the relation.

$$E^2 = \frac{\text{Between-class variation}}{\text{Total variation}} \tag{41}$$

For our example we substitute the information from Table 60 into formula (8) of Chapter 22 to get the between-class variation thus,

$$\begin{aligned}\frac{\text{Between-class}}{\text{variation}} &= \frac{(156.7)^2}{8} + \frac{(115.9)^2}{8} + \frac{(62.5)^2}{5} + \frac{(40.7)^2}{3} \\ &+ \frac{(36.0)^2}{3} + \frac{(5.8)^2}{1} + \frac{(34.7)^2}{3} - \frac{(452.3)^2}{31} = 349.68\end{aligned}$$

We have already computed the total variation on page 413 and found it to be 929.83. Substituting in (41) we get

$$E^2 = \frac{349.68}{929.83} = .3760$$

$$E = .614$$

This value is slightly higher than the absolute value of the linear correlation coefficient, $r = .515$. Is the difference between the two great enough to signify that there is a difference between the corresponding universe parameters, η and ρ? If so, it would mean that there is a "significant" departure from linearity in the association between migration and natural increase. This, in turn, would mean that the assumption of linearity is not justified in analyzing the relationship between the two characteristics because it *underestimates* the degree of the association and *incorrectly* describes the nature of the association since it assumes the wrong sort of form for the regression equation.

To answer the question the difference between the squares of the two observed coefficients is tested rather than the difference between the two coefficients themselves, that is, we test to see if $E^2 - r^2$ is significantly different from zero. The test may be made by an analysis of variance.

Let us first set up a regular analysis of a variance table based upon the results of the computations above. Table 61 shows the form of such an analysis, and Table 62 shows, the actual analysis for this example. From Table 62 we see that we cannot reject the hypothesis that $\eta = 0$.[17]

---

[17] Unlike r (and ρ), r_c (and ρ_c) and E (and η) can take only positive values. In nonlinear regression and correlation the *direction* of association does not necessarily remain constant throughout the range of observed X values; therefore, we cannot in general use the concept of direction of association in curvilinear correlation.

Table 61. ANALYSIS OF VARIANCE OF N VARYING UNITS IN CHARACTERISTIC Y FOR TESTING SIGNIFICANCE OF CORRELATION RATIO BASED ON MEANS OF m CLASS INTERVALS OF CHARACTERISTIC X

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	Ratio of variances, F
Total (Units about $\bar{Y}$).....	Σy^2	$N-1$		
Column means about $\bar{Y}$..........	$E^2 \Sigma y^2$	$m-1$	$\dfrac{E^2 \Sigma y^2}{m-1}$	$\dfrac{E^2(N-m)}{(1-E^2)(m-1)}$
Units about column means......	$(1-E^2)\Sigma y^2$	$N-m$	$\dfrac{(1-E^2)\Sigma y^2}{N-m}$	

Table 62. ANALYSIS OF VARIANCE OF 31 ECONOMIC AREAS IN PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING (Y) FOR TESTING SIGNIFICANCE OF CORRELATION RATIO BASED ON MEANS OF 7 CLASS INTERVALS [a] OF NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X)

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total (economic areas about $\bar{Y}$)......	929.83	30		
Column means about $\bar{Y}$............	349.61	6	58.27	2.41
Economic areas within columns about column means.................	580.21	24	24.18	

$P[F_{6.24} = 2.41] > .05$

[a]. We could consider that we had eight class intervals in Table 60, but since one of them includes no cases and we make no use of the fact that the intervals are of equal width, it is more logical to ignore the interval that contains no cases. Thus, we have only seven independent observations and six degrees of freedom for column means about $\bar{Y}$.

Source: Table 60.

Since this is true, there is no real point in testing the significance of the difference between η^2 and ρ^2; however, we will make this test in order to demonstrate the method. Table 63 gives the general form of such a test, and Table 64 makes the test for our particular problem. Since the F in Table 64 is less than one, we cannot reject the hypothesis that $\eta^2 - \rho^2 = 0$. This indicates that we have no significant departure from linearity.

This test for linearity as well as the test for the significance of E can be shortened, just as the F test for the significance of r was shortened.

Table 63. ANALYSIS OF VARIANCE OF N VARYING UNITS IN CHARACTERISTIC Y FOR TESTING SIGNIFICANCE OF DIFFERENCE BETWEEN THE SQUARE OF THE CORRELATION RATIO BASED ON MEANS OF m CLASS INTERVALS OF CHARACTERISTIC X AND THE SQUARE OF THE CORRELATION COEFFICIENT BASED ON A LINEAR REGRESSION ON X

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	Ratio of variances, F
Total (Units about $\bar{Y}$)....	Σy^2	$N - 1$		
Regression line about $\bar{Y}$.........	$r^2 \Sigma y^2$	1		
Column means about regression line.....	$(E^2 - r^2)\Sigma y^2$	$m - 2$	$\dfrac{(E^2 - r^2)\Sigma y^2}{m - 2}$	$\dfrac{(E^2 - r^2)(N - m)}{(1 - E^2)(m - 2)}$
Units about column means............	$(1 - E^2)\Sigma y^2$	$N - m$	$\dfrac{(1 - E^2)\Sigma y^2}{N - m}$	

Table 64. ANALYSIS OF VARIANCE OF 31 ECONOMIC AREAS IN PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING FOR TESTING SIGNIFICANCE OF DIFFERENCE BETWEEN THE SQUARE OF THE CORRELATION RATIO BASED ON MEANS OF 7 CLASS INTERVALS OF NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X) AND THE SQUARE OF THE CORRELATION COEFFICIENT BASED ON A LINEAR REGRESSION ON X

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total..........................	929.83	30		
Regression line about $\bar{Y}$.............	246.614	1		
Column means about regression line..	19.317	5	3.863	.160
Units within columns.............	580.214	24	24.18	

$P[F_{5.24} = .160] > .05$

Source: Data from Table 55; computations from Tables 58, 62.

Since the several amounts of variation shown in the sum of squares columns in all three pairs of tables—59 and 58, 61 and 62, 63 and 64—all contain the term Σy^2, and since F is always a ratio of variances formed from these amounts of variation, Σy^2 always cancels out in the expression

for F. Therefore, we can determine F's without actually dividing up the variation as in a full analysis of variance, simply from the values of r, E, N, and m. Such expressions have already been shown in Tables 57, 61, and 63, but they will be summarized here along with the type of hypothesis each tests.

SUMMARY OF F TESTS FOR THREE TYPES OF HYPOTHESES

Type of hypothesis tested		Formula for F	Degrees of freedom
I	$\rho^2 = 0$	$F = \dfrac{r^2(N-2)}{1-r^2}$	$n_1 = 1$ $n_2 = N - 2$
II	$\eta^2 = 0$	$F = \dfrac{E^2(N-m)}{(1-E^2)(m-1)}$	$n_1 = m - 1$ $n_2 = N - m$
III	$\eta^2 - \rho^2 = 0$	$F = \dfrac{(E^2-r^2)(N-m)}{(1-E^2)(m-2)}$	$n_1 = m - 2$ $n_2 = N - m$

The formulas for F given above are the ones we usually evaluate directly in a practical situation without going through a complete analysis of variance.

Estimation of η. Since the third test listed above failed to show the required value for F, we conclude there is no significant departure from linearity in this example. Hence, there is no practical need of making an estimate of the universe parameter, η, corresponding to the observed E. To illustrate the procedure, however, we shall proceed and make the estimate as if we were going to use it. The formula for estimating η is as follows,

$$\hat{\eta}^2 = \frac{E^2(N-1) - (m-1)}{N-m} \qquad (42)\ {}^{18}$$

It can be seen that this formula takes into account the number of degrees of freedom sacrificed by making the hypothetical regression line go through the column means. Substituting the values for our example in (42), we have

$$\hat{\eta}^2 = \frac{(.614)^2(31-1) - (7-1)}{31-7} = .2212$$

$$\hat{\eta} = .470$$

---

[18] Formula (42) is a special case of the more general formula,

$$\hat{\rho}^2_{YX} = 1 - (1 - \rho^2)\frac{(N-1)}{(N-m)}$$

$$= \frac{\rho^2(N-1) - (m-1)}{N-m}$$

which is used by some writers for the estimation of universe values of ρ^2, ρ_c^2, and η^2 where N is the number of cases, m is the number of constants determined for the regression equation, and ρ is the observed coefficient of correlation whether it is linear, curvilinear, or a correlation ratio.

The fact that $\hat{\eta}$ is lower than $\hat{\rho}$ confirms the test for nonlinearity and means that there is no indication that these 31 observations came from a universe where the form of association is nonlinear.

The coefficient $\hat{\eta}$ as defined by formula (42) is identical with the coefficient ϵ as defined in Chapter 22 by the relation,

$$\epsilon^2 = 1 - \frac{V_w}{V_t} \tag{43}$$

If we write formula (42) thus,

$$1 - \epsilon^2 = \frac{V_w}{V_t}$$

and compare this with formula (40) for E^2, we see that the difference between the two is that $1 - E^2$ is equal to the ratio of unexplained *variation* (sum of squares) to total *variation*, while $1 - \epsilon^2$ is equal to the ratio of the unexplained *variance* to total *variance*. Thus, ϵ takes into account the degrees of freedom on which the coefficient is based whereas E does not.

To illustrate the interpretation of a different verdict from the test of hypothesis III above, let us suppose that E^2 had been much larger than r^2 and that consequently the third test had given a very large F, beyond the value required for .001 level of significance. Then we should have rejected the hypothesis $\eta^2 - \rho^2 = 0$ and concluded that the association in the universe was nonlinear. In such a case we would have made an estimate of η from formula (42), and this estimate would have been higher than our estimate of ρ, since the estimate of degree of association on the assumption of linearity in a universe where association is nonlinear is always an *under*estimate. Therefore, if the purpose of a correlation analysis is merely to establish the *existence* of association in the universe and if r is significantly different from zero, no incorrect conclusion will result from omitting a test for linearity. If, however, the test above of hypothesis I (or some equivalent test) shows that r is *not* significantly different from zero, there is no justification for concluding that there is no association in the universe unless hypothesis II is tested. For nonlinear associations, while often not differing greatly in form from linear, may differ so greatly as to have an r of zero even when E is very high.

OTHER TYPES OF HYPOTHESES RELATING TO CORRELATION COEFFICIENTS, THEIR APPROPRIATE TESTS, AND THEIR APPLICATION

There are a number of types of hypotheses relating to correlation coefficients which one may wish to test. In addition to the three types listed on page 455, we have already in setting up confidence limits implied

the test of another type of hypothesis, which we shall now consider explicitly. Let us continue the sequence of Roman numerals begun in the summary on page 455 to designate the types of hypotheses to be tested and the types of tests appropriate for testing them.

Type IV, $\rho = A$. Type IV is a hypothesis that in the universe the coefficient of correlation is equal to any specified value, A (type I is a special case of this). In symbols we can express this hypothesis thus, $\rho = A$ or $\rho - A = 0$. As in the setting up of confidence limits, the Z transformation is used for testing hypotheses of this type. Let us illustrate the test by inquiring whether our observed r of $-.515$ could have come from a universe where $\rho = -.80$. We shall follow the regular five steps we have ordinarily used in testing a hypothesis.

1. *Formulation of the hypothesis to be tested.* This is already done when A is selected and in this example is

$$\rho = -.80 \text{ or } \rho - (-.80) = 0$$

Since we know, however, that r's based on 31 observations from a universe where $\rho = -.80$ will not be normally distributed, we shall transform our hypothesis to an equivalent hypothesis concerning Z_u, because we know that sample Z's are approximately normally distributed. By referring to Appendix Table G, we find that a Z_u of 1.098613 corresponds to a $|\rho|$ of .80. Therefore, we shall test the hypothesis that $Z_u = 1.098613$.

2. *Description of the sampling distribution of the statistic.* In a universe with a Z_u of 1.098613, the sampling distribution of Z's computed from 31 observations will have its mean at 1.098613, will be approximately normal, and will have a standard deviation equal to

$$\sigma_Z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{28}} = .18896$$

3. *Determination of the probability that a Z as unusual as that observed would be expected.* First we must transform our observed r into a Z. By use of Appendix Table G we find that a Z of .569511 corresponds to our observed r of $-.515$. We now express the observed Z as a deviation from the mean of the sampling distribution in terms of standard deviation units, thus

$$\frac{Z - Z_u}{\sigma_Z} = \frac{.569511 - 1.098610}{.11896} = 2.75$$

When we refer the value 2.75 to Appendix Table C, we discover that the probability of so unusual a deviation is .006.

4. *Rejection of the hypothesis.* Since P is so small, we reject the actually tested hypothesis, $Z_u = 1.098613$ which means that we can also reject the equivalent hypothesis,

$$\rho = -.80.$$

5. *Interpretation of the results.* We are justified in concluding that the universe from which the observations on the 31 economic areas may be considered a random sample does not have an association between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators for the year 1945 of a degree so great as that measured by $-.80$.

Limited use of such tests. A test such as the one just made is of little use in practical sociological research situations, for we rarely have a theory formulated so precisely as to suggest the universe parameter of correlation. In the example above it is as if there were actually some theory that 64 percent (r^2) of the variation in percentage net change due to migration is associated with variation in average annual rate of natural increase. At later stages in the development of sociology such tests may be needed more often, but at present situations calling for them are rare. Furthermore, as we have indicated previously, the setting up of confidence limits implicity tests many such hypotheses simultaneously, for the two confidence limits indicate two values beyond which any value tested for the universe would get a verdict of rejection at the level of significance specified. The 99-percent confidence limits of the estimated value of ρ determined on pages 427–429, $r_1 = -.0819$ and $r_2 = -.784$, tell us that for any universe value between these two the verdict of a test such as that just made will be nonrejection of the hypothesis, and that for any universe value beyond these two the verdict will be rejection at the one-percent level of significance. Therefore, it was needless to test the hypothesis that $\rho = A$ when we knew confidence limits in advance, for they indicate what range of universe values deduced through theory would be "acceptable" as far as this one set of observations can give an answer.

Type V, $\rho_1 - \rho_2 = 0$. The type of test we are much more likely to need than the above is the type explained in Chapter 19 for one characteristic—that is, a test relating to observations made on the same characteristics from two samples. Continuing our sequence of Roman numerals, we shall call this type of hypothesis and its test type V. The hypothesis can be expressed in symbols as $\rho_1 - \rho_2 = 0$, or $\rho_1 = \rho_2$.

For illustration of this type of hypothesis and its test, let us assume once more that the 115 white tenant farm women of the Tobacco Piedmont (on whom data are available for two characteristics) are a random sample of all white tenant farm women in the area, and similarly that the 119 white tenant farm women of the Deep South (on whom data are available for the same two characteristics) are a random sample of all white tenant farm women in that area. In the investigation of the association between fertility (Y) as measured by the number of children ever borne, and

education (X) as measured by the number of grades of school completed, the following results were obtained:

Piedmont group	Deep South group
$N = 115$	$N = 119$
$r = -.210$	$r = -.387$

In both groups the women with more education tend to have fewer children, but in the Deep South group the tendency is more marked. Is it enough higher to lead us to conclude that there must be differences in the cultures of the two groups which make education a more decisive factor in the determination of the number of children a Deep South tenant farm woman bears than in the case of the Piedmont woman? If we find that the answer is yes, this would suggest an inquiry into the differences in content and methods of education of the two groups and also into the pattern of attitudes toward childbearing in the two groups. Conceivably the association between the two factors may be lower in the Piedmont group either because the type of education received is different or because the pattern of high fertility is more deeply set and less affected by education.

Before trying to plan an inquiry to throw light on the reasons for the difference, however, we should like to be assured of the reliability of the difference we have observed, for we must remember that the coefficients of correlation we are using as measures of the degree of association between the two characteristics are subject to sampling error and so is the difference between two of them. The precise question we must phrase is this: On the basis of the difference observed between the two r's, can we safely infer that a difference in the same direction exists between the two universe ρ's? Of course, we know that no absolute assurance is possible from statistical analysis of observed data, but we might define the expression "safely infer" to mean that we follow a procedure which will lead us to correct conclusions in 95 or 99 times out of 100. Following the research leads suggested by differences in correlation coefficients takes both time and money, and it would be wasteful to undertake an elaborate study to inquire into the reasons for a difference if in the universes there is really no difference.

Therefore, we test the significance of the difference between the two observed r's in the following manner.

1. *Formulation of hypothesis.* Since r_1 is smaller than r_2, what we wish to establish is that $\rho_1 - \rho_2 < 0$. All alternatives to this hypothesis are included in the general null hypothesis, $\rho_1 - \rho_2 \geq 0$, of which the limiting case which would yield a P of greatest value when tested by our observations is $\rho_1 - \rho_2 = 0$, which we use as our specific null hypothesis

to be tested. Again we shall use Fisher's Z transformation and actually test not this hypothesis but a hypothesis equivalent to it, $Z_{u_1} - Z_{u_2} = 0$.

2. *Description of the sampling distribution of the statistic.* The sampling distribution of $Z_1 - Z_2$ in samples drawn from the universe described by the null hypothesis is approximately normal with a mean of zero and a standard deviation of

$$\sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \qquad (44)$$

where N_1 = number of observations in one sample
and N_2 = number of observations in the other sample
Substituting the N's of our example in (44), we find,

$$\sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{115 - 3} + \frac{1}{119 - 3}}$$
$$= .132474$$

3. *Determination of the probability that a statistic as unusual as the one observed would be expected.* First by using Appendix Table G we transform our sample r's into Z's and find that $Z_1 = .213590$ and $Z_2 = .408032$. Then the statistic whose sampling distribution is described above is

$$Z_1 - Z_2 = .213171 - .408267 = -.195096$$

Next we express the observed difference as a deviation from the mean of its sampling distribution in standard deviation units,

$$\frac{(Z_1 - Z_2) - 0}{\sigma_{Z_1 - Z_2}} = \frac{-.195096}{.132474} = -1.47$$

Referring this value to Appendix Table C, we find that $P = .1416$.

4. *Nonrejection of the hypothesis.* Since a deviation as unusual as this would be expected 14 out of 100 times from samples of the specified sizes drawn from the universe described by the null hypothesis, we cannot reject the hypothesis $Z_{u_1} = Z_{u_2}$ or its equivalent, $\rho_1 = \rho_2$.

5. *Interpretation of the test.* We have not been able to show that the observed difference between the two correlation coefficients signifies any real difference between the two universe coefficients. Let us emphasize again, however, that we have *not* proved or affirmed in any way the null hypothesis $\rho_1 = \rho_2$. All we have done is to show that it is one (of many) possible situations in the universes from which we might have gotten the sort of results we did in our sample. In fact, because our P is only .14, we may be somewhat suspicious of the hypothesis, even though we cannot reject it.

Inverse use of Test V to determine number of cases necessary to establish the fact of a difference in the universe. Let us suppose, for the moment, that there is other evidence that the pattern of high fertility among tenant farm women in the Piedmont is more firmly set and less easily affected by education than in the Deep South. How can we use the above information to plan an inquiry which will provide a basis for detecting a significant difference in the degree of association between the two factors in the different areas? The following procedure is suggested, although it is only a roughly approximate device. We have seen that the sampling error of $Z_1 - Z_2$ is dependent only on the numbers of cases in the samples. Therefore, even if there were a universe difference, $Z_{u_1} - Z_{u_2}$, of approximately the size we observed, that is, $-.19$, it would be impossible to establish the existence of this difference from samples of size 116 and 119. Let us assume that there is such a difference in the universe and determine how large our samples would have to be to give convincing evidence of its existence. We use the same relations used in testing the hypothesis, but we proceed in reverse direction.

First, we must decide what level of significance we demand to consider evidence "convincing." Suppose we choose the one-percent level. The number of sigma units corresponding to a P of .01 is approximately 2.6 (from Appendix Table C). Then in the relation,

$$\frac{Z_1 - Z_2}{\sigma_{Z_1 - Z_2}} = \text{sigma units required for significance}$$

we can substitute the difference we are assuming to exist in the universe for $Z_1 - Z_2$ and -2.6 as the number of sigma units required. Thus,

$$\frac{-.19}{\sigma_{Z_1 - Z_2}} = -2.6$$

and solving for $\sigma_{Z_1 - Z_2}$, we have

$$\sigma_{Z_1 - Z_2} = \frac{.19}{2.6} = .07 \text{ (approximately)}$$

Let us further suppose that we are planning for samples to be of the same size. Then we can substitute N for N_1 and N_2 in the formula (44),

$$\sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \tag{44}$$

and also the required value of $\sigma_{Z_1 - Z_2}$, obtaining

$$.07 = \sqrt{\frac{1}{N - 3} + \frac{1}{N - 3}} = \sqrt{\frac{2}{N - 3}}$$

If we neglect the 3, which is small in comparison to N, we can easily solve this equation for N by squaring and proceeding with simple algebra, thus,

$$.0049 = \frac{2}{N}$$

$$N = \frac{2}{.005} \text{ (approximately)}$$

$$N = 400$$

We see that we should need two samples of approximately 400 each to demonstrate that the coefficient of correlation in one universe is higher than that in the other, if there is a difference in Z_v's of $-.19$ and *if* we observed a difference in sample Z's of about this size. Since we are usually not at all sure of the validity of these two assumptions, it is better to take a sample somewhat larger than that indicated—say 500 in this case. It must be remembered that this procedure is not based upon rigorous statistical theory, but is offered simply as a guide of a very approximate nature.

Even so, it is urged that students contemplating correlation analysis of data from field studies being planned estimate from similar work of others, or even guess (if necessary) what the values of coefficients of correlation they wish to compare will be. Then through some such process as the above they should see if the size of samples they are planning is great enough to enable them to recognize such a difference as significant. It is evident that correlation analysis made for the purpose of establishing the significance of a difference in degree of association between two characteristics in two samples requires a relatively great number of cases unless the coefficients are very high.

Applicability of Test V to comparison of coefficients of correlation observed between characteristics distributed among the same sample. We may wish not only to compare two coefficients of correlation between the same two characteristics observed in different samples, but also to compare two coefficients of correlation between different pairs of characteristics observed in the same sample. For instance, let us suppose that in the Deep South group of 119 tenant farm women the coefficient of correlation between number of children ever borne and annual family cash income is $-.60$. It would be interesting to know if this r is significantly higher than the r of $-.39$ observed between number of children ever borne and education of mother—that is, if variation in fertility is more closely associated with variation in present income than with education of mother. The test of type V just explained is *not* appropriate for such a situation; it is appropriate only when the two r's are from different, independent

samples. In the present illustration, any test of the difference between the two r's would have to take into account the fact that the two characteristics whose associations with fertility are measured by the two r's may themselves be associated. For such a situation there is no straightforward test such as the one just given. There is a formula for the standard error of the difference between two r's—for instance, r_{YX} and r_{YZ}—which involves also r_{XZ}, but the form of its sampling distribution is not known. Under certain circumstance we use a test of type V with qualifications explained below. Some light can be thrown on the interrelationship of three variables by the methods of multiple and partial correlation analysis to be presented in Chapter 25.

Applicability of test of type V to demographic problems. It may be noted that to illustrate tests of type V we chose a problem involving groups of individuals rather than demographic areas as varying units. One may well ask, is this type of test ever appropriate for testing the differences between coefficients of correlation observed for demographic characteristics distributed among demographic units? Let us examine several of the more common situations where the need for such a test arises and see if the situations are appropriate for the above test.

Situation 1: The most common situation is where the varying demographic units are the areal subdivisions of a larger demographic unit or population (for instance, the 48 states of the United States). In this situation there is no practical sampling situation, for all units in a limited universe are measured. Therefore, one cannot compare a coefficient of correlation observed between two demographic characteristics with any other coefficient between the same two characteristics observed on a set of units of the same order for the United States at the same time, because there is no other set. It is possible that one may wish to make two other sorts of comparisons, however.

The first sort of comparison is that of an r computed for two characteristics distributed among the demographic units of one population with an r for the same characteristics distributed among the demographic units of another population. For instance, if for some group of states other than North Carolina, South Carolina, and Georgia, we compute the coefficient of correlation between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators in 1945, we may wish to compare it with the r of $-.515$ in our example. In this case the test of type V is appropriate.

If we consider the states as populations and their counties as the varying demographic units, we often have occasion to use such tests. If the two states are contiguous or if the counties of one state are in any way affected by the counties of the other, the situation does not theoretically meet the criterion of *independence* of samples. However, as explained in Chapter 20

the contiguity of the counties within a state has already violated the criterion of independence for the varying units of the population. If that camel has been swallowed, one should not strain at this gnat. As suggested before, careful investigation is needed into the whole matter of the effects of geographical contiguity of varying units on their distributions of characteristics and the associations between them. Until the results of such investigation appear, however, population students will continue to determine coefficients of correlation between characteristics as distributed among the counties of different states, the census tracts of different cities, perhaps the states of different regions, and in such cases the test of type V seems to be applicable for testing the significance of the difference between coefficients of correlation between the same characteristics for varying units of the same order of different populations.

The second type of comparison is that of an r computed for all varying units of a population with an r between the same two characteristics computed for the same varying units at *different times*. For instance, suppose that after the release of the 1950 census we find that the coefficient of correlation between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators in 1950 in the 31 economic areas used in our example is $-.84$. We should certainly be interested in comparing this r with our r of $-.515$ for 1945 to see if there has been a change in the degree of association of the two characteristics. The applicability of the above test of type V depends on whether the 31 observations for 1945 may be considered independent of the 31 observations for 1950. It is evident that they cannot, for the measures for any one economic area in 1945 would be related to the measures for that economic area in 1950. The relation could be described by computing the correlation coefficient between the two dates for either or both measures. If these correlations are different from zero, a test of type V, which requires r's from independent samples would not be appropriate.

If the correlations between the two dates for measures on the same characteristic are positive—that is, if the economic areas with large numbers of nonwhite farm operators per 100 white farm operators in 1945 tended to have large numbers of nonwhite farm operators per 100 white farm operators in 1950—then the test of type V would result in fewer errors of the first kind and in more errors of the second kind.[19] This means that the standard error of the difference of the Z's is overestimated by formula (44), that we are likely to get the verdict "insignificant" when there is a real difference more often than when there is no positive correlation between the two dates. Sometimes using such a test is termed

---

[19] See page 323.

"being more conservative," but whether or not this is true depends on what one is trying to prove. However, the use of a test of type V in such a situation with a .01 level of significance is the equivalent of using a more rigid level of significance (requiring a P of smaller numerical value) but we do not know *how much* more rigid. Thus, if in the hypothetical example, we find that the measures for the two decades are positively correlated (and this is reasonable to suppose), we use a test of type V and get a verdict that the difference is significant, we can know that the (unknown) correct test would also show the difference to be significant. If, however, we get a verdict that the difference is not significant, we have no way of knowing whether the (unknown) correct test would have shown it to be significant or not. With these limitations on the interpretation of such a test it is all right to use the type V test in such a situation, or in the situation described on pages 462–463. A word of caution is necessary, however, as to the necessity of observing that the correlation between the measures of each characteristic for the two decades is *positive*. If it is negative, then a test of type V will underestimate the standard error of the difference between the Z's; thus, it will give more errors of the first kind and fewer errors of the second kind than a correct test. Its use is rarely justified in this case.

Situation 2: Another common situation in demographic research occurs when a population is divided into two or more groups of units according to some criterion. For instance, one might divide the population of the United States into predominantly rural counties and predominantly urban counties. Or one might divide the population into urban and rural, and then subdivide the urban into several groups of varying units (cities) by size of city. Unless the criterion for division makes the parts correspond to some areal units, the situation becomes more complicated. (For instance, if we divide the population of the United States by the criterion of working age into two groups of people, those of working age and those not of working age, there is no series of areal units or demographic entities, arbitrary or not, which contain only people of one group.)

The most common sort of situation here is probably the comparison of coefficients of correlation between characteristics as observed in groups of cities of the same size class. (Each group may be composed of all the cities of the same size class or of a sample of them.) Suppose one has obtained a coefficient of correlation between sex ratio and percentage of females married for each of two city size groups—cities under 100,000 and cities of 100,000 or over. In such a case a type V test might be used but with a recognition that the order of varying units in one group is hardly equivalent to the order of varying units of the other, since they are differentiated on the basis of size, and that this may confuse the interpretation of the test. If there are more than two such groupings, then the methods of

analysis of covariance are applicable. An illustration of this type of problem will be presented in Chapter 24.

RANK CORRELATION

We indicated in the classification of characteristics by their measures of incidence that for a type II B characteristic data are given with respect to the degree of incidence in the form of rankings of the individual units. Such data are often obtained from subjective judgments as to some characteristic, in which cases they may be quite as valid as data secured by objective devices, if the person doing the ranking is both expert and impartial. Nevertheless, one cannot be sure that another judge would rank the units in exactly the same way; therefore, we continually strive to develop more objective measuring techniques.

Sometimes we take units for which we have quantitative measures and rank them on the basis of these quantitative measures either to simplify handling the date (with resultant loss of information) or to minimize the effects of a few very extreme cases.[20]

We will discuss two coefficients of rank correlation. The first is the Spearman rank correlation, the more familiar of the two. The second is Kendall's τ, a relatively recent development that has certain advantages over Spearman's rank correlation.

We shall illustrate the methods of computing the rank correlation coefficients with a hypothetical example. Suppose that 10 couples participating in programs of the Cooperative Agricultural Extension Service are ranked at the end of a year as to their excellence in farm management and in home management during the year. Let us suppose that the farm management rank is assigned by the County Agricultural Extension Agent who has observed the farmers' practices throughout the year and that the home management rank is assigned by the County Home Demonstration Agent, who has observed the wives' practices. The hypothetical data are given in columns (1) and (2) of Table 65. From inspection one can detect that there is a positive association between excellence in farm and home management, even though only one couple, Couple H, received the identical rank in both.

Since ranks are not literally measures of incidence, the methods of correlation analysis given earlier in this chapter are not applicable for the analysis and description of the association between excellence in farm and home management as distributed among these 10 couples. Instead, as a measure of the degree of association between the two characteristics, we must employ some measure of rank correlation.

---

[20] See Maurice G. Kendall, *Rank Correlation Methods* (London: Charles Griffin, 1948), pp. 14–15.

Table 65. COMPUTATION TABLE FOR SPEARMAN'S RANK CORRELATION COEFFICIENT FROM HYPOTHETICAL RANKINGS OF 10 COUPLES ON FARM AND HOME MANAGEMENT

Couple	Rank in farm management (1)	Rank in home management (2)	Difference in ranks, D (3)	D^2 (4)
A	7	5	2	4
B	4	2	2	4
C	2	1	1	1
D	9	8	1	1
E	1	4	-3	9
F	6	3	3	9
G	3	6	-3	9
H	10	10	0	0
I	5	7	-2	4
J	8	9	-1	1
Sums:			0	42

Formula	*Evaluation*
$r_r = 1 - \dfrac{6\Sigma D^2}{N(N^2 - 1)}$	$r_r = 1 - \dfrac{6(42)}{10[(10)^2 - 1]}$ $= .74$

Source: Hypothetical data.

The Spearman rank correlation. The Spearman rank correlation, which we shall designate by r_r, is defined by the formula

$$r_r = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \tag{45}$$

where N = number of paired observations

D = difference between two ranks of one varying unit

It does not matter which of the two characteristics is listed first for the computation of the D's, since the sum of the D's will be zero in either case, and we shall use only the sum of the D^2's, in which their signs disappear. Columns (3) and (4) of Table 65 show the simple computations necessary to obtain the expression $\Sigma D^2 = 42$. Substitution in (45) is shown in the lower part of Table 65 with

$$r_r = .74$$

Spearman's rank correlation coefficient r_r can be interpreted as being the equivalent of Pearson's product moment coefficient r *only if* the differences in the ranks on each characteristic represent equal differences

in degree of incidence; for instance, if the difference in degree of excellence in farm management between farmers with ranks 10 and 9 is equal to the differences in degree of excellence between farmers with ranks 9 and 8, 8 and 7, 7 and 6, and so on. This is not generally the case, for if the distribution of the characteristic comes anywhere near to approaching normality, the differences in measures represented by a difference of one in the middle ranks will be much smaller than the difference in measures represented by a difference of one in the extreme ranks. If it can be assumed that the characteristics are normally distributed, one can obtain from tables the value of r corresponding to a given observed value of r_r.[21] Since the corresponding r will be slightly higher and since one seldom has enough information to justify the assumption of normality in the distribution of characteristics of sociological interest measured by ranks, it seems advisable not to use such a table. Therefore, it is better to interpret r_r as a rough measure of degree of association, having the same range of possible values as r, but not exactly equivalent to r and therefore not subject to the more elaborate methods developed for product moment correlation analysis.

Sampling distribution of the Spearman rank correlation coefficient. The sampling distribution of the Spearman rank correlation coefficient has not been satisfactorily described except in the cases where $\rho_r = 0$ and N is large and where N is 8 or less.[22] When N is equal to 8 or less, exact tests of the null typothesis that $\rho_r = 0$ are possible by the use of Olds' Tables,[23] but we rarely work with so few cases. An approximate test of r_r for N between 9 and 20 can be made using the t distribution discussed in Chapter 16. In this case we compute t as follows:

$$t = r_r \sqrt{\frac{N - 2}{1 - r_r^2}} \qquad (46)\ ^{24}$$

Using this value of t and $(N - 2)$ degrees of freedom we can utilize Appendix Table D to determine the approximate probability of observing such an r_r as we observed in a sample of N cases from a universe in which $\rho_r = 0$. For our case,

$$t = .74 \sqrt{\frac{10 - 2}{1 - (.74)^2}} = 13.1$$

---

[21] Such tables can be found in almost any text on educational statistics or in R. E. Chaddock, *Principles and Methods of Statistics* (Boston: Houghton Mifflin, 1925), Appendix E, p. 464.

[22] Kendall, *op. cit.*, pp. 46–49.

[23] E. G. Olds, "Distributions of Sums of Square of Rank Differences for Small Numbers of Individuals," *Annals of Mathematical Statistics*, XIV (1943), pp. 149–152. Similar tables are reproduced in Kendall, *op. cit.* Appendix Table 2.

[24] *Ibid.*, p. 48.

The probability of such a value of t is less than .001, so we can say that there is a significant positive association between farm and home management. However, when $\rho_r \neq 0$ the sampling distribution of r_r is unknown, and, therefore, we are unable to set any confidence limits on r_r.

When N is greater than 20 (though this is a doubtful point), we can test the significance of r_r by making use of the fact that

$$\sigma_{r_r} = \sqrt{\frac{1}{N-1}} \qquad (47)$$

We should keep in mind that the tests of the significance of the Spearman rank correlation coefficient are approximations except when N equals 8 or less.

Kendall's coefficient of rank correlation, τ (tau). One method of computing τ is first to arrange the cases in rank order from lowest rank to highest rank according to one of the variables.[25] (We are using lowest rank here to mean lowest numerical rank; thus, 1 is lower in rank than 5, 6 is lower in rank than 7, etc. We refer to 10 as the highest rank in our illustration even though it may represent the poorest farm or home management.) In Table 66 we have reordered the cases of our illustration from lowest to highest rank in farm management as is shown in column (1). From column (2), the rank in home management, we then compute a measure P. Each case contributes a certain number of points toward P as is shown in column (3). The contribution of a particular case to P is the number of cases listed *below* it in the table that have higher rankings in both farm management and home management. Since the cases are arranged in order of rank in farm management, all cases listed below any given case have higher rankings in farm management. Therefore, for each case we merely count the number of cases listed below it in the table that have rankings in home management higher in numerical value than the case under consideration. For example, the contribution of case H is zero because there are no cases listed below it. The contribution of case D is 1, as shown in column (3), because the case listed below it has a higher numerical rank. The contribution of case J is also 1 because of the two cases listed below it, only one of them, H, has a higher numerical rank. The contribution of case A is 3 since all three cases listed below it have higher numerical ranks. Skipping to case G we see that it contributes 4 points to P because of the seven cases listed below it only four of them have higher numerical ranks. In this manner we obtain the entries in column (3). P is the sum of the entries in column (3).

Having computed P, we can now proceed to compute τ_s. (Tau is

---

[25] Kendall also gives methods for computing τ without arranging the cases in order of one of the rankings. *Ibid.*, p. 6.

Table 66. COMPUTATION TABLE FOR KENDALL'S τ FROM
HYPOTHETICAL RANKINGS OF 10 COUPLES ON FARM
AND HOME MANAGEMENT

Couple	Rank in farm management (1)	Rank in home management (2)	Contribution to P (3)
E	1	4	6
C	2	1	8
G	3	6	4
B	4	2	6
I	5	7	3
F	6	3	4
A	7	5	3
J	8	9	1
D	9	8	1
H	10	10	0
			$P = 36$

Formula		*Evaluation*
$\tau_s = \dfrac{2P}{\frac{1}{2}N(N-1)} - 1$	$=$	$\dfrac{2(36)}{\frac{1}{2}(10)(10-1)} - 1$
	$=$	$.47$

Source: Hypothetical data.

given the subscript s in this case in order to indicate that it is the rank correlation in a sample rather than in the universe from which the sample was drawn.) The formula for τ_s is

$$\tau_s = \frac{2P}{\frac{1}{2}N(N-1)} - 1 \qquad (48)$$

Taking $P = 36$ from Table 66 and substituting in formula (48) we get

$$\tau_s = \frac{2(36)}{\frac{1}{2}(10)(10-1)} - 1 = .47$$

Sampling distribution of Kendall's τ. The advantage which this measure of rank correlation has over Spearman's rank correlation is that the sampling distribution is known. The sampling distribution of Kendall's τ converges to normal very rapidly and can be considered normal

whenever N is equal to or greater than 10. The sampling distribution for N less than 10 has been computed and is recorded in tables.[26] Actually instead of testing the significance of τ, we test the significance of another measure which we can derive from P. This measure is

$$S = 2P - \frac{1}{2}N(N - 1) \tag{49}$$

The standard error of S is given by the formula

$$\sigma_s = \sqrt{\frac{1}{18}N(N - 1)(2N + 5)} \tag{50}$$

For our illustrative problem

$$S = 2(36) - \frac{1}{2}(10)(10 - 1) = 27$$

$$\sigma_s = \sqrt{\frac{1}{18}(10)(10 - 1)(20 + 5)} = 11.2$$

To test the significance of S we form the quotient

$$\frac{S}{\sigma_s} = \frac{27}{11.2} = 2.4$$

From Appendix Table C we see that the probability of getting such an S if our sample of 10 cases were drawn from a universe in which farm and home management are unrelated is .016. Thus, we see that the relationship is significant at the .05 level.

To be entirely correct we should correct S for continuity before making the above test. This correction involves subtracting 1 from S if S is positive and adding 1 if S is negative. Making the correction for continuity in the above illustration does not affect our conclusion that the relationship is significant at the .05 level of significance.

If there are numerous tied rankings in one or both of the variables, then the computation of τ may require special handling.[27] However, such treatment is beyond the range of this text. Kendall also gives methods for setting up approximate confidence limits on τ and methods of getting

---

[26] *Ibid.*, Appendix Table 1.
[27] *Ibid.*, Chapter 3.

partial rank order correlations. The student is referred to the original volume for these methods.[28]

Uses of rank correlation coefficients. Rank correlation coefficients can be used not only on data originally obtained as ranks, but also on data originally obtained as measures, which are later ranked. For instance, the 48 states might be ranked according to percentage net change due to migration and according to average annual rate of natural increase for the decade 1940–1950 and a rank correlation coefficient used to describe the association. The disadvantages of using a rank correlation coefficient in such a case are first, that it does not use all the information of the observations, being based simply on relative order of a measure rather than on its numerical value; and second, the precise testing of hypotheses about the universe is difficult or impossible in some situations. Yet, because of the simplicity of computing a rank order correlation, one is often used in the two following cases: (1) when the series of units is so short that no reliable measure can be obtained anyway and when a rank order correlation is sufficient for a rough descriptive measure of the degree of association; and (2) when one wishes to make a hasty or preliminary description of degree of association to ascertain more accurately than by inspection if the degree is great enough to warrant the computation of a more elaborate coefficient.

SUGGESTED READINGS

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chaps. 22 and 23.

Ezekiel, Mordecai, *Methods of Correlation Analysis*, 2d ed. (New York: Wiley, 1941), Chaps. 3–8.

Kendall, Maurice G., *Rank Correlation Methods* (London: Charles Griffin, 1948).

McNemar, Quinn, *Psychological Statistics* (New York: Wiley, 1949), Chaps. 6, 7, and 8.

Mode, Elmer B., *The Elements of Statistics* (New York: Prentice-Hall, 1941), Chap. 12.

Peatman, John Gray, *Descriptive and Sampling Statistics* (New York: Harper, 1947), Chaps. 9 and 10.

Snedecor, George W., *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), Chap. 7.

Yule, G. Udny, and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), Chaps. 9, 10, 11, 13, and 14.

---

[28] *Ibid.*, Chapters 4 and 8.

〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜〜

Analysis of Covariance

NATURE AND UTILITY OF THE METHODS OF
ANALYSIS OF COVARIANCE

Overlapping of bodies of method. In dividing statistics of relationship into the bodies of method designated by the titles of the chapters in Part IV, we do not mean to suggest that the groups of procedures are isolated and unrelated. The use of analysis of variance in problems of correlation has already been illustrated, as well as the use of correlation in problems of contingency and analysis of variance. In the methods of this chapter, however, there is even more overlapping, for analysis of covariance is essentially a synthesis of the methods of correlation and of analysis of variance to provide for analysis of associations more complex than can be investigated by either correlation or analysis of variance alone, using these terms in a narrow sense. Obviously, then, an understanding of the two chapters immediately preceding this one is prerequisite to learning the methods of analysis of covariance.

Situations to which analysis of covariance is applicable. In the outline of statistics of relationship given on page 348 we have seen that analysis of covariance is designed for investigating associations between two or more quantitative characteristics and one or more nonquantitative characteristics. The simplest case is that where there are only two quantitative characteristics and one nonquantitative characteristic. This is treated in the present chapter. In fact we cannot treat more than two quantitative characteristics simultaneously until we have developed the methods of multiple and partial correlation and regression; the restrictions of equal or proportionate class frequencies along with the other requirements considered in the chapter on analysis of variance limit severely the application of analysis of covariance in problems in which there is more than one nonquantitative classification.[1]

---

[1] See references listed in footnote 10, p. 397.

Suppose for a series of units we have measures of variables X and Y, and also have the units grouped into m classes with respect to a nonquantitative characteristic, A. The methods of correlation provide the techniques for investigating the total association between X and Y, and the methods of analysis of variance for investigating the total association between X and A, or Y and A, but neither of these provides the techniques for investigating the partial association between two of the three characteristics while the third is allowed for (as the methods of partial correlation "hold constant" one of the three variables while the association between the other two is investigated). This is the chief use of analysis of covariance, and by its methods we can investigate the partial association between the two quantitative variables, X and Y, with differences in A taken into account, and also the partial association between A and either one of the variables, with the other variable "held constant." It may seem logically correct to use the term "analysis of covariance" for only the first case, that is, for the analysis of the covariance of two variables, but it is customary to use the term in a broader sense to apply to both cases. In fact, it is the latter case that is generally referred to when the term is used, and this is the case investigated in the basic analysis of a covariance table.

Questions answered by analysis of covariance. It is possible to anticipate the type of questions about the interassociations between X, Y, and A which analysis of covariance will enable us to answer. We shall regard Y as the "dependent" variable as in the preceding chapter. Some of the questions are as follows:

1. Is there a significant partial association between Y and A when differences with respect to X are allowed for? (Are there significant differences between the set of m means in Y for the classes of A after they have been "adjusted" for differences in X?)

2. If the answer to the first question is yes, what is the nature of the partial association between Y and A when differences with respect to X are allowed for? (What are the adjusted Y means of the m classes?)

3. Is there a significant partial association between X and Y when differences with respect to A are allowed for? (Is there a significant "average within-class" regression?)

4. Do the individual associations between X and Y in the m classes of A differ significantly in nature? (Are the b_{YX}'s of the m classes significantly different?)

5. Is there a significant association between the class means in X and Y?

6. If the answer to the preceding question is yes, is the nature of the "between-class" association significantly different from the nature of the "average within-class" association?

Not all of the above questions will be pertinent or meaningful in every analysis of a covariance problem. If the chief interest is in the association

between Y and A (corresponding to "yield" and "treatments" in certain agricultural problems, or to "achievement score" and "methods of teaching" in educational problems), the primary interest is usually in questions 1 and 2. If the chief interest is in the association between X and Y, the primary interest is usually in questions 3 and 4. If the chief interest is in the joint association of all three characteristics, the interest is in all of the questions, including 5 and 6. Of course, all of the questions may be repeated with X and Y interchanged if the problem is such that X may also be considered the "dependent" variable.

By means of an illustrative problem involving population data we shall present the procedures for answering the above questions in the order listed. Preliminary to presenting the methods of analysis of covariance, however, we shall use correlation and analysis of variance to investigate the existence and nature of the three total associations involved—the associations between X and Y (total correlation), between X and A, and between Y and A (analysis of variance).

The problem. The problem consists of investigating for 38 selected metropolitan areas having over 250,000 population the associations between sex ratio of the population aged 15–24, percent of females single (never married), and regional location.[2] More specifically, the characteristics whose associations are to be investigated for 38 metropolitan areas with over 250,000 population are the following:

$Y =$ percent of females 14 years of age and over that were single (never married) in 1950. For brevity we shall refer to this as the percent of females single.

$X =$ number of males between 15 and 24 per 100 females between 15 and 24 in 1950. We shall refer to this measure as the sex ratio.

$A =$ a classification of the United States into four regions. This classification corresponds to Howard W. Odum's regional classification[3] except that we have combined the Southwest, the Northwest, and the Far West into the West in order to get sufficient cases in one classification. The names of the regions and the number of metropolitan areas considered in each are as follows:

The Northeast	12 metropolitan areas
The Southeast	8 metropolitan areas
The Middle States	9 metropolitan areas
The West	9 metropolitan areas

The data for the problem are shown in Table 67.

---

[2] The metropolitan areas included were selected from those for which data were available at the time of writing. Selection was made in order to emphasize certain aspects of analysis of covariance, and because of this selection the conclusions of the illustrative problem should not be accepted without qualifications.

[3] Howard W. Odum and Harry E. Moore, *American Regionalism* (New York: Holt, 1938), Chaps. XVII and XVIII.

Table 67. PERCENT OF FEMALES 14 YEARS OF AGE AND OVER SINGLE AND
SEX RATIO OF POPULATION AGED 15–24 FOR 38 SELECTED METROPOLITAN
AREAS WITH OVER 250,000 POPULATION, GROUPED BY REGIONS, 1950

Region (A) and metropolitan area	Percent of females 14 and over single Y	Sex ratio of population aged 15–24 X	Region (A) and metropolitan area	Percent of females 14 and over single Y	Sex ratio of population aged 15–24 X
Northeast			Richmond......	22	80
Albany.........	22	109			
Baltimore......	21	94	*Middle States*		
Buffalo.........	24	88	Akron...........	18	86
Harrisburg.....	20	110	Chicago........	20	88
Johnstown......	25	90	Cincinnati......	21	89
Pittsburgh......	23	90	Cleveland......	19	79
Providence.....	26	100	Dayton........	20	86
Rochester......	22	91	Detroit........	18	88
Springfield—			Kansas City....	16	83
Holyoke......	27	79	Milwaukee.....	23	91
Syracuse.......	23	103	Wheeling—		
Washington,			Steubenville...	21	89
D. C.........	22	90	*West*		
Worcester......	25	80	Dallas.........	15	92
Southeast			Denver........	20	77
Atlanta........	18	84	Houston.......	14	94
Birmingham....	17	84	Los Angeles.....	15	96
Louisville.......	18	89	Portland.......	16	88
Memphis.......	15	78	San Antonio....	19	86
Miami.........	14	92	San Diego......	15	102
Nashville.......	19	84	San Francisco...	16	87
New Orleans....	20	85	Seattle.........	16	90

Source: 1950 Census of Population. Preliminary Reports. Series PC-5.

INVESTIGATION OF TOTAL ASSOCIATIONS

Association between Y and X. Let us first consider the data on the
percent of females single and the sex ratio without regard to the regional
classification. By computation from the data of Table 67 the following
sums are obtained:

$$N = 38 \qquad\qquad \Sigma XY = 66{,}496$$
$$\Sigma Y = 745 \qquad\qquad \Sigma X = 3{,}391$$
$$\Sigma Y^2 = 15{,}061 \qquad\qquad \Sigma X^2 = 304{,}845$$

By substitution in formulas (1), (2), and (3) of Chapter 23, the following intermediate measures may be obtained:

$$\Sigma y^2 = 15,061 - \frac{(745)^2}{38} = 455$$

$$\Sigma x^2 = 304,845 - \frac{(3,391)^2}{38} = 2,243$$

$$\Sigma xy = 66,496 - \frac{(3,391)(745)}{38} = 15$$

Substitution of these values in formulas (4), (8), and (9) of Chapter 23 gives

$$r_{YX} = \frac{15}{\sqrt{(455)(2,243)}} = .015$$

$$b_{YX} = \frac{15}{2,246} = .0067$$

$$a_{YX} = \frac{745 - (.0066)(3,391)}{38} = 19.0$$

These last two values may be combined into a regression equation, thus,

$$Y_c = 19.0 + .0067X \tag{1}$$

Figure 42 shows the scatter plot for the 38 metropolitan areas in the percent of females single and the sex ratio with the fitted regression equation (1).

While it is fairly obvious that an r of .015 with only 38 cases is not significant, we shall test the significance of it using analysis of variance in order to illustrate the test. We shall develop formulas for obtaining the "explained" sum of squares and the "unexplained" sum of squares slightly different from those used in the preceding chapter.

It will be remembered that r^2 is the proportion of variation in Y which we consider "explained" by the regression on X, that is,

$$\text{Explained sum of squares} = r^2 \Sigma y^2 \tag{2}$$

From the equation,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \tag{3}$$

let us substitute the right side for r in equation (2), obtaining

$$\text{Explained sum of squares} = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} \Sigma y^2 = \frac{(\Sigma xy)^2}{\Sigma x^2} \tag{4}$$

Figure 42. Regression of Percent of Females Single on Sex Ratio. (Source: Table 67.)

Since the unexplained variation is equal to the total variation minus the explained variation, we have

$$\text{Unexplained sum of squares} = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} \tag{5}$$

Evaluating (4) and (5) with the data of our problem, we obtain

$$\text{Explained sum of squares} = \frac{(15)^2}{2,243} = 0.10$$

$$\text{Unexplained sum of squares} = 455 - 0.1 = 454.9$$

These two sums of squares together with the total sum of squares can now be entered into an analysis of the variance table as shown in Table 68.

Table 68. ANALYSIS OF VARIANCE OF 38 SELECTED METROPOLITAN AREAS IN PERCENT OF FEMALES SINGLE (Y) FOR TESTING SIGNIFICANCE OF RE-GRESSION ON SEX RATIO (X), 1950

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total (areas about $\bar{Y}$).........	455	37		
Regression line about $\bar{Y}$........	0.1	1	0.1	.01
Areas about regression line.....	454.9	36	12.6	

Source: Table 67.

Since the resulting F is less than one there is no significant association between X and Y. It must be remembered that the association just described is a *total* association between X and Y, and the investigation of the total association has not taken into account in any way the regional classification, A.

Association between Y and A. Now we shall investigate the association between the percent of females single and the regional classification shown in Table 67. By obtaining the regional sums in Table 67 and substituting in formula (8) of Chapter 22, we get the between-region sum of squares as follows:

$$\frac{\text{Between-region}}{\text{sum of squares}} = \frac{(280)^2}{12} + \frac{(143)^2}{8} + \frac{(176)^2}{9} + \frac{(146)^2}{9} - \frac{(745)^2}{38} = 294$$

Using this and the total sum of squares $(\Sigma y^2 = 455)$, we make the analysis of variance shown in Table 69. The association between the percent of females single and the regional classifications is highly significant as indicated by the F obtained from the mean square variances. The degree of association as measured by the unbiased correlation ratio can be found by substituting the mean square variances in formula (7) of Chapter 22.

Table 69. ANALYSIS OF VARIANCE OF 38 SELECTED METROPOLITAN AREAS IN PERCENT OF FEMALES SINGLE (Y) BY REGIONAL CLASSIFICATION (A)

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total (areas around $\bar{Y}$).......	455	37	12.30	
Between regions (region means about $\bar{Y}$)..........	294	3	98.00	20.68
Within regions (areas about region means)............	161	34	4.74	

$P[F_{3.34} = 20.68] < .001$

Source: Table 67.

$$\epsilon^2 = 1 - \frac{V_w}{V_t}$$

$$\epsilon^2 = 1 - \frac{4.74}{12.30} = .6146$$

$$\epsilon = .784$$

Since the classes of A are unordered, the direction of association can have no meaning. The nature of association is best described by computing the four class means in percent of females single, which will be presented shortly.

Association between X and A. By exactly the same procedures used in investigating total association between Y and A we can investigate association between X and A. The segregation of the total sum of squares $(\Sigma x^2 = 2,243)$ into the portions arising from differences between region means and from differences among metropolitan areas within regions is shown in Table 70. The F indicates that the set of region means in sex ratio differ significantly at the .05 level of significance. The degree of association is again measured by ϵ,

$$\epsilon^2 = 1 - \frac{51.62}{60.62} = .1485$$

$$\epsilon = .385$$

Again the direction of association can have no meaning. The nature of both total associations, between Y and A and between X and A is described in Table 71.

Table 70. ANALYSIS OF VARIANCE OF 38 SELECTED METROPOLITAN AREAS IN SEX RATIO OF POPULATION AGED 15–24 (X) BY REGIONAL CLASSIFICATION (A)

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total (areas around $\bar{X}$)......	2,243	37	60.62	
Between regions (region means about $\bar{X}$).........	488	3	162.67	3.15
Within regions (areas about region means)...........	1,755	34	51.62	

$P[F_{3.34} = 3.15] < .05$

Source: Table 67.

Table 71. REGION MEANS IN PERCENT OF FEMALES SINGLE (Y) AND SEX RATIO OF POPULATION AGED 15–24 (X), 38 SELECTED METROPOLITAN AREAS, 1950

Region	Mean of measures for metropolitan areas in a region	
	Percent of females single	Sex ratio of population aged 15–24
All regions................	19.6	89.2
Northeast................	23.3	93.7
Southeast................	17.9	84.5
Middle States.............	19.6	86.6
West....................	16.2	90.2

Source: Table 67.

In summary, we have investigated the three total associations and found two of them significant. We did not find any significant total association between X and Y. In order to examine the possible effects of one association on another, we turn to the methods of analysis of covariance.

COMPUTATIONS FOR ANALYSIS OF COVARIANCE

For the basic analysis of covariance we need one additional computation. We have divided the total variation of X into two parts, between

regions and within regions, and we have done the same for the total variation of Y. We must now do this for the covariation of X and Y. We have already computed the total covariation, $\Sigma xy = 15$. To get the between-region covariation we use the following general formula:

$$\begin{aligned}\text{Between-class}\\\text{covariation}\end{aligned} = \frac{(X._1)(Y._1)}{k_1} + \frac{(X._2)(Y._2)}{k_2} + \cdots + \frac{(X._i)(Y._i)}{k_i} + \cdots$$
$$+ \frac{(X._m)(Y._m)}{k_m} - \frac{(X..)(Y..)}{N} \tag{6}$$

where $X._i$ = the sum of the X's in the ith class
$\quad\quad Y._i$ = the sum of the Y's in the ith class
$\quad\quad k_i$ = the number of cases in the ith class
$\quad\quad X..$ = the sum of all the X's
$\quad\quad Y..$ = the sum of all the Y's

One can see the similarity between this formula and formula (8) of Chapter 22. Computing regional sums from Table 67 and substituting in formula (6) gives

$$\begin{aligned}\text{Between-region}\\\text{covariation}\end{aligned} = \frac{(1124)(280)}{12} + \frac{(676)(143)}{8} + \frac{(779)(176)}{9} + \frac{(812)(146)}{9}$$
$$- \frac{(3391)(745)}{38} = 235$$

The fact that this portion of the total covariation is greater than the total covariation need not be disturbing when one remembers that covariation, unlike variation, can be either positive or negative. To find the within-region covariation we subtract the result above from the total and get

$$\begin{aligned}\text{Within-region}\\\text{covariation}\end{aligned} = 15 - 235 = -220$$

Partial association of X and Y (Question 3). With these data and the data from Tables 70 and 71 we make Table 72. Table 72 answers question 3 on page 474. We observe here that the within-region correlation between X and Y is $-.414$. This within-region correlation is the correlation between X and Y after the regional variations have been "removed." The fact that the total correlation is effectively zero while the within-region correlation is $-.414$ indicates that the percent of females single is related to the sex ratio but that the relationship was being obscured by the regional differences. (We will shortly make a test to discover if the nature of the relationship is the same for all regions.)

One of the major uses of analysis of covariance is to examine relationships of this sort, the relationship between X and Y when the effects of A are removed. The total correlation can be zero and the within-class correlation can be significantly different from zero, as in this case. It is possible for the total correlation to be significantly different from zero and the within-class correlation to be zero or of an opposite sign from the total

Table 72. COMPUTATION OF WITHIN-REGION CORRELATION BETWEEN PER-CENT OF FEMALES SINGLE (Y) AND SEX RATIO OF POPULATION AGED 15–24 (X), 38 SELECTED METROPOLITAN AREAS, 1950

| Source of variation | Sums of squares and products | | | | Correlations | | |
	Degrees of freedom	Σy^2	Σx^2	Σxy	$\dfrac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}$ r^2	r	Degrees of freedom
Total...........	37	455	2,243	15	.00022	.015	36
Between region...	3	294	488	235			
Within region....	34	161	1,755	−220	.17129	−.414	33

Source: Tables 67, 69, and 70.

correlation. And, of course, the total correlation may not be significantly different from the within-class correlation.

Partial association of Y and A (Question 1). It is possible to construct a table slightly different from Table 72 in order to answer question 1—is there a significant relationship between the percent of females single and regions after the differences in sex ratio have been allowed for? Table 73, known as the basic analysis of covariance table, answers this question.

In this table the "unexplained sums of squares" are also called the "sums of squares of errors of estimate." The first entry under "Unexplained sums of squares" is easy to understand, for it is a quantity we have dealt with previously. In the chapter on correlation we have seen that this quantity is the sum of the squares of the deviations of the observed Y's from the line of total regression of Y on X. It is this quantity which we shall divide into two portions to make the basic analysis of covariance F test.

The first portion, 133.5, is the part of the within-region sum of squares unexplained by a regression, which we call "the average within-region" regression. In Table 74 we shall employ a procedure for averaging a number of individual class regressions. The procedure involves adding for each region the Σx^2 measured from its own region mean $\bar{X}_i$, the Σy^2 measured from its region mean $\bar{Y}_i$, and the Σxy also measured from the region means. Then by the ordinary formulas for b and r the regression and correlation

coefficients are computed from these composite Σx^2, Σy^2, and Σxy. The correlation coefficient thus computed is the within-region correlation, computed from a slightly different approach in Table 72. The squares of deviations from this "average within-region" regression are the unexplained sums of squares within regions. If regressions from several groups may be averaged, this is the way it is done, and the average within-class regression is considered to represent the nature of the partial association of Y and X within classes of A.

Table 73. ANALYSIS OF COVARIANCE OF 38 SELECTED METROPOLITAN AREAS IN PERCENT OF FEMALES SINGLE (Y) AND SEX RATIO OF POPULATION AGED 15–24 (X) BY 4 REGIONS (A)

Source of variation	Sums of squares and products				Errors of estimate		
	Degrees of freedom	Σy^2	Σx^2	Σxy	Unexplained sums of squares $\Sigma y^2 - \dfrac{(\Sigma xy)^2}{\Sigma x^2}$	Degrees of freedom	Mean square variance
Total..........	37	455	2,243	15	454.9	36	
Between region...	3	294	488	235			
Within region....	34	161	1,755	−220	133.5	33	4.0
For testing adjusted region means					321.4	3	107.1

$$F_{3,33} = \frac{107.1}{4.0} = 26.8$$

$$P[F_{3,33} = 26.8] < .001$$

Source: Table 72.

We shall speak more of the interpretation of the average within-class regression presently. Our interest in it for interpreting Table 73 is not for itself but for its use in "adjusting" [4] the region means in Y $(\bar{Y}_i)$. We have said that we wish to "hold constant" or "allow for" region differences in X in order to test the significance of the partial association of Y and A. The average within-class regression enables us to do this by the process of "adjusting" the $\bar{Y}_i$'s, as we shall find shortly. When the $\bar{Y}_i$'s are adjusted for differences in X, will the set of them still differ significantly among

---

[4] There is an average within-region correlation coefficient, r, and an average within-region regression coefficient, b, but there is no one constant of a regression equation, a, which along with b would specify a single line representing the average within-region regression. The constant a to be used with the average within-region regression coefficient, b, is different for each region. We use the regression line described by this value of a along with the average within-region regression coefficient b in adjusting the region means.

themselves? The highly significant F of Table 73 answers the question affirmatively. The sum of squares labeled "For testing adjusted region means" is a function of the differences between the adjusted means. The difference between the sum of squares unexplained by the total regression and the sum of squares unexplained by the within-region regression arises from differences in $\bar{Y}_i$'s not accounted for by differences in $\bar{X}_i$'s. This point may clear up somewhat after we have actually adjusted the region means, which we shall do shortly.

The sum of squares, 321.5, in Table 73 is obtained by subtracting the within-region unexplained sum of squares from the total unexplained sum of squares. The degrees of freedom for errors of estimate are one less than the original degrees of freedom for the total and within-class unexplained sums of squares because the regression line in each case uses up an additional degree of freedom. The degrees of freedom for testing adjusted region means are obtained by subtraction since the unexplained sum of squares is obtained by subtraction in this case. The F is formed by dividing the mean square variance for testing adjusted region means by the mean square variance within regions. In this case,

$$F_{3,33} = \frac{107.2}{4.0} = 26.8$$

By comparison with Table 68 we see that the regional differences in the percent of females single are more important after adjustment for sex ratio differences than they were when the sex ratios were ignored. Frequently regional differences will show up as significant only after adjustments have been made for the effects of some other variable.

Additional computations in analysis of covariance. In order actually to adjust the region means in the percent of females married for region differences in sex ratio and in order to answer all of the questions which analysis of covariance is designed to answer, we must make certain additional computations. We will illustrate the formal computations before proceeding to interpret the results. For each class (region) we will need the following data: k (number of units in the class), ΣY, ΣY^2, ΣX, ΣX^2, ΣXY. These data must be obtained for each of the m classes of A, and when they are obtained for each class, they comprise all the information necessary for the computations. For a full analysis of covariance there will be needed 27 columns in a computation table and as many rows as there are classes in A (m) plus 4. Table 74 shows such a table with the row and column designations.

Columns (1)–(13). Column (1) contains the designations of the rows. The entries in its first m rows are the numbers or names of the m classes of A. Its next four rows contain the designations "Sums," "Total," "Between class," and "Within class." In the following columns the processes

Table 74. COMPUTATION TABLE FOR INVESTIGATING TOTAL ASSOCIATIONS AND FOR INVESTIGATING PARTIAL ASSOCIATIONS BETWEEN Y AND A DIFFERENCES IN A ALLOWED FOR BY ANALYSIS OF

Class	k_i	ΣY	$\dfrac{(\Sigma Y)^2}{k_i}$	ΣY^2	Σy^2	ΣX	$\dfrac{(\Sigma X)^2}{k_i}$	ΣX^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	*12*	*280*	6,533	*6,582*	49	*1,124*	105,281	*106,392*
2	*8*	*143*	2,556	*2,603*	47	*676*	57,122	*57,262*
3	*9*	*176*	3,442	*3,476*	34	*779*	67,427	*67,533*
4	*9*	*146*	2,369	*2,400*	31	*812*	73,260	*73,658*
Sums	38	745	14,899	15,061	161	3,391	303,090	304,845
Total	38	745	14,606	15,061	455	3,391	302,602	304,845
Between class					294			
Within class					161			

Class	$\Sigma x^2 \Sigma y^2$	$\sqrt{\Sigma x^2 \Sigma y^2}$	$r = \dfrac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$	$\bar{X}_i = \dfrac{\Sigma X}{k_i}$	$x = \bar{X}_i - \bar{X}$
(1)	(17)	(18)	(19)	(20)	(21)
1	54,439	233.3	−.596	93.7	4.5
2	6,580	81.12	−.333	84.5	−4.7
3	3,604	60.03	.633	86.6	−2.6
4	12,338	111.1	−.828	90.2	1.0
Sums					
Total	1,020,565	1010	.015	89.2	
Between class	143,472	378.8	.620		
Within class	282,555	531.6	−.414		

of obtaining the entries are the same for the four rows corresponding to the four regions, but they vary somewhat for these last four rows, and separate instructions will be given for them. As can be seen, not every column has an entry for these four rows.

The entries of columns (2), (3), (5), (7), (9) and (11) for the first four rows are italicized. This is done to show that they are filled in with the data secured from Table 67—the number of observations (k_i) and the five sums, ΣY, ΣY^2, ΣX, ΣX^2, and ΣXY for each class. The first step after

BETWEEN Y, X, AND A BY CORRELATION AND ANALYSIS OF VARIANCE
WITH DIFFERENCES IN X ALLOWED FOR AND BETWEEN X AND Y WITH
COVARIANCE, 38 SELECTED METROPOLITAN AREAS, 1950

Class	Σx^2	ΣXY	$\dfrac{(\Sigma X)(\Sigma Y)}{k_i}$	Σxy	$b = \dfrac{\Sigma xy}{\Sigma x^2}$	Explained $\dfrac{(\Sigma xy)^2}{\Sigma x^2}$	Unexplained $\Sigma y^2 - \dfrac{(\Sigma xy)^2}{\Sigma x^2}$
(1)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1	1,111	26,088	26,227	−139	−.125	17.4	31.6
2	140	12,056	12,083	−27	−.193	5.2	41.8
3	106	15,272	15,234	38	.358	13.6	20.4
4	398	13,080	13,172	−92	−.231	21.3	9.7
Sums	1,755	66,496	66,716	−220			103.5
Total	2,243	66,496	66,481	15	.0067	.10	454.9
Between class	488			235	.482	113.3	180.7
Within class	1,755			−220	−.125	27.5	133.5

Class	(within class b)x	$\hat{Y}_i = \dfrac{\Sigma Y}{k_i}$	$\hat{Y}_i - bx$	$b\Sigma X$	$\Sigma Y - b\Sigma X$	$a = \dfrac{\Sigma Y - b\Sigma X}{k_i}$
(1)	(22)	(23)	(24)	(25)	(26)	(27)
1	−.562	23.3	23.9	−140.5	420.5	35.0
2	.588	17.9	17.3	−130.5	273.5	34.2
3	.325	19.6	19.3	278.9	−102.9	−11.4
4	−.125	16.2	16.3	−187.6	333.6	37.1
Sums						
Total		19.6		22.72	722.3	19.0
Between class				1,634	−889	−23.4
Within class				−423.9	1,169	30.8

Source: Table 67.

entering these sums in the table is to add the corresponding sums for
the four classes and enter the additions under the appropriate column in
two rows—the "Sums" row and the "Total" row. For instance, when all of
the entries in column (2) are added, their sum, 38, is placed in column (2)
opposite "Sums" and opposite "Totals." Similarly when the entries in
column (3) (ΣY) are added, their sum, 745, is placed in the same two rows
of column (3), and so on for columns (5), (7), (9), and (11). In columns

(2), (3), (5), (7), (9), and (11) there are no entries opposite the last two rows of the table, "Between class" and "Within class."

The first four rows and the "Total" row of these first 13 columns are simply the computations for obtaining the familiar Σy^2, Σx^2, and Σxy for each class and for the total series. For these rows the entries in column (6) are obtained by subtracting the entries in column (4) from the entries in column (5). The entries in column (8) are subtracted from those in column (9) to get the entries in column (10), and the entries in column (13) are obtained by subtracting the entries in column (12) from those in column (11).

For further work we shall be chiefly concerned with these entries in columns (6), (10), and (13), and it is only in these columns that we shall need to make entries for the remaining three rows.

For the "Sums" row the entries for the four classes are added for columns (6), (10), and (13). By considering the nature of the operations performed so far, one can see that the "within-class" variation in Y and X and the "within-class" covariation in Y and X have been computed directly by summing the variation and the covariation of the classes. Therefore, the entries in columns (6), (10), and (13), opposite "Sums" are written again in the same column opposite "Within class." Finally, from the relation used in analysis of variance, the entries for columns (6), (10), and (13) opposite "Between class" are obtained by subtracting the entry opposite "Within class" from that opposite "Total." For instance, in column (6), 161 is subtracted from 455 to get 294, the "Between class" Σy^2.

Columns (14), (15), and (16). The entries for all rows except the row labeled "Sums" are obtained similarly for these columns. Note that the entries for column (15) are the products of the entries in columns (13) and (14). The computations for columns (14) and (16) follow directly from the column headings.

It is from column (16) that the quantities for the basic analysis of covariance table (Table 73) are obtained. This table answers the question regarding relationship between Y and A when differences in X are accounted for. If the examination of this relationship is the primary purpose of the investigation and the answer at this point is "no relationship," there will be no need of proceeding further.

Columns (17), (18), and (19). These three columns are for the purpose of obtaining seven coefficients of correlation, one for each row of the computation table except the "Sums" row. The entries for all seven rows are made similarly. It is from column (19) that we would get the data for Table 72 had we not already done these computations. Notice, however, that from Table 72 there is no hint that the direction of relationship within one of the classes is opposite to the direction of relationship for the other classes. This fact, also shown in column (14), will raise questions, to

be discussed later, regarding the meaningfulness of the "within-class correlation."

Columns (20), (21), (22), (23), and (24). These columns relate to the class means and the total mean (grand mean), and they are not filled in for the "Sums," "Between-class," or "Within-class" rows. In fact, only columns (20) and (23) are filled in for the "Total" row. For column (22) the entry in column (21) for each class is multiplied by the "Within-class" regression coefficient, that is the entry in column (14) in the "Within-class" row, $-.125$ in our example. The entries in column (24) are the "adjusted" class means mentioned on page 484. They will be discussed further after the explanation of the computation table is completed.

Columns (25), (26), and (27). These columns are for the purpose of obtaining seven a's to use with the b's of column (14) to form seven regression equations corresponding to the seven coefficients of correlation in column (19). No entries are made in these columns for the "Sums" row. The entries are made similarly for the four classes and for the "Total," "Between-class," and "Within-class" rows, except that for the three latter rows the "Total" row's entries for ΣX and ΣY are used, although the b's corresponding to each row are used.

INTERPRETATION OF RESULTS FROM THE COMPUTATION TABLE

The computation table just completed (Table 74) provides the data needed for making Tables 68–73, had we not made these tables previously. In addition to providing the data for making these basic tables, Table 74 provides additional information needed to interpret properly Tables 72 and 73.

The test made in Table 73 indicated that there were significant differences between the regional means in the percent of females single even after regional differences in the sex ratio were allowed for. In other words, regional differences in the sex ratio do not account for the regional differences in the percent of females single. Column (23) of the computation table shows the region means in the percent of females single. Column (24) shows these means adjusted for regional differences in the sex ratio. The process of adjusting can perhaps be made clearer by a chart. In Figure 43 look first at the solid dots representing the observed means and the solid line representing the average within-region regression. (We are calling it the average within-region regression here, although we shall see later that it is not from this line that the deviations are measured whose sum of squares forms the within-region unexplained sum of squares.) The point marked with a cross mark is the point determined by the means of the X and Y distributions ($\bar{X} = 89.2$, $\bar{Y} = 19.6$). Now imagine that each dot is shifted parallel to the average within-class regression line until it lies

Figure 43. Observed Region Means in Percent of Females Single (Y) and Region Means Adjusted for Average Within-region Regressions on Sex Ratio (X). (Source: Table 74.)

directly over the point on the X axis representing $\bar{X}$, 89.2. This shift is indicated by the dotted lines. We have now "adjusted" all of the points until their X value is equal to $\bar{X}$. Their resulting Y values are the adjusted region means shown in column (24) of the computation table.

If the dots were originally all very close to the solid line, the adjusting process would have the effect of evening out their differences in Y. This would be the case if differences in sex ratio actually accounted for differences in region means in the percent of females single. In our case we have seen that significant differences still remain. Columns (23) and (24)

of Table 74 show that the rank order of the regions in the percent of females single does not change with the adjustment for differences in sex ratio. Clearly there are other factors making for the regional differences in the percent of females single. The nature of association of Y and A when differences in X are allowed for (question 2, page 474) is described by columns (23) and (24) of Table 74.

The average within-region regression (Question 3). We have already used the average within-region regression to adjust the region means. Now we must use it to test the significance of the partial correlation between X and Y when variations in A are accounted for. This is the within-class correlation computed in Table 72 and again in Column (19) of Table 74, and it is the correlation coefficient associated with the average within class regression. Table 75 shows the usual analysis of variance test of sig-

Table 75. ANALYSIS OF THE WITHIN-REGION VARIANCE IN PERCENT OF FEMALES SINGLE (Y) FOR TESTING SIGNIFICANCE OF THE AVERAGE WITHIN-REGION REGRESSION ON SEX RATIO (X), 38 SELECTED METROPOLITAN AREAS, 1950

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Metropolitan areas within regions.....	161	34		
Average within-region regression lines about $\bar{Y}$........................	27.5	1	27.5	6.79
Metropolitan areas around average regression lines......................	133.5	33	4.05	

$P[F_{1.33} = 6.79] < .05$

Source: Table 74.

nificance. The sums of squares are taken from the "Within-class" row of columns (6), (15), and (16) of Table 74. (The necessary sums of squares may also be gotten from the last line of Table 73.) Thus, from Table 75, we see that our within-class correlation of $-.414$ is significant at the .05 level of significance. The fact that our total correlation between X and Y is not significant and the within-class correlation is significant suggests that in the total association the heterogeneity of regions is in some way obscuring the association actually existing between Y and X within the regions. Table 75, then, answers question 3, page 474 affirmatively.

We have spoken apparently inconsistently of the average within-region regression line, at times as if it were one line and at times as if it were more than one line. We have shown a line in Figure 43 supposedly repre-

Figure 44. Individual Region Regressions of Percent of Females Single (Y) on Sex Ration (X) and Average Within-region Regressions. (Source: Table 74.)

senting the average within-region regression and then have said that it was not from this line that the deviations of the metropolitan areas were measured to give the sum of squares unexplained by the average within-region regression. Let us clear up the confusion. It will be remembered that the Σx^2, Σy^2, and Σxy computed for each region from the means of

that region were added for the four regions and entered first in the "Sums" row and then in the "Within-class" row. From these "Within-class sums" the average within-class regression coefficient, $b = -.125$, and the average within-class correlation coefficient, $r = -.414$, were computed and also the sums of squares explained and unexplained by the regression. Now the deviations whose squares compose this last sum are for any region the distance from the observed Y to a line which has a $b = -.125$ but which passes through $\bar{Y}_i$ and $\bar{X}_i$ of that region. Figure 44 shows the average within-region regression line for each region as well as the individual regression line for each region, the two crossing at the point $(\bar{X}_i, \bar{Y}_i)$ for each region. If, as in Figure 43, we need for the whole series some *one* line to represent the average within-region regression, we choose the line with a $b = -.125$ which passes through the point $(\bar{X}, \bar{Y})$. This is the proper line to use for adjusting the subregion means, but one must keep in mind the fact that it is not the line from which the squared Y deviations of the 38 metropolitan areas equals 133.5. In fact, there can be no such line, for the line of total regression, fitted by the least squares criterion, is the line from which the sum of the squared Y deviations of the 38 metropolitan areas is the smallest, and we have seen that this sum is 454.9. Any line with a sum of squares of deviations from it smaller than 454.9 is impossible, and, therefore, there is no one average within-region regression line, although there is an average within-region regression coefficient and also a correlation coefficient.

The individual subregion regressions (Question 4). Columns (14) and (19) of Table 74 show r's and b's for each region. Since one of the regions, the Middle States, shows a positive relationship between X and Y while the others show negative relationships, we ask the question, are the individual regressions significantly different? Table 76 is designed to answer this question. The first and last sums of squares in Table 76 are taken from the "Sums" and the "Within-class" rows of column (16) of the computation table, and the sum of squares arising from differences between the individual region regressions is obtained by subtraction. The number of degrees of freedom for the first sum of squares is $N - m - 1$, as obtained in Table 75. The degrees of freedom for the second sum of squares is $m - 1$. The degrees of freedom for the deviations from individual region regressions can be obtained by subtraction, or are $N - 2m$ since two degrees of freedom are lost for the regression of each of the m classes. The F test of Table 76 indicates that the individual regressions are not significantly different from each other even though one of the lines has a positive slope and the others have negative slopes. However, the probability of obtaining such an F is almost .05, and, since .05 is an entirely arbitrary level of significance, some question should be raised in our minds as to whether or not these regressions are significantly different.

If the regressions are significantly different, then we had no right to

average the four regressions as if they were really from the same universe subject only to sampling fluctuations. And if we cannot average them to get an average within-region regression, then we cannot obtain a within-region coefficient of correlation or do the basic analysis of covariance shown in Table 73. We can describe the relationship between X and Y within each region separately, but a conservative interpretation would be to forego the composite description obtained for the partial association. However, since the probability obtained in Table 76 is greater than .05, we will assume that we were justified in averaging the regressions to obtain an average within-class regression. Because of its importance to the

Table 76. ANALYSIS OF ERRORS OF ESTIMATE FROM AVERAGE WITHIN-REGION REGRESSION

Source of variation	Sums of squares	Degrees of freedom	Mean square variance	F
Deviations from average within-region regressions....................	133.5	33	4.05	
Differences between individual region regressions....................	30.0	3	10.00	
				2.90
Deviations from individual region regressions....................	103.5	30	3.45	

$P[F_{3,30} = 2.90] > .05$

Source: Table 74.

interpretation of tables such as Tables 72, 73, and 75, the test of the significance of the difference of the individual regressions shown in Table 76 should always be made.

If we had found the individual region regressions to be significantly different, we would have wanted to test the significance of each separately. Table 77 shows these tests. For each region F is obtained by dividing the explained mean square variance, shown in column one, by the mean square variance of errors of estimate, shown in column four. From Table 77 we see that two of the individual region regressions are not significant, and we note that one of these is the regression within the Middle States. The two individual region regressions that are significant are both negative; so we have additional evidence for accepting our average within-region regression, which is also negative.

Association between Y and X for class means (Question 5). The sums of squares already computed are sufficient to answer the question of

whether or not there is a significant association between Y and X for the region means. Table 78 shows the customary test, although it is hardly necessary since such a small amount of the variation is explained by the regression. The fact that the between-region regression is not significant while the average within-region regression is significant indicates that re-

Table 77. ANALYSIS OF VARIANCE OF 38 SELECTED METROPOLITAN AREAS FOR TESTING SIGNIFICANCE OF THE INDIVIDUAL REGION REGRESSIONS OF PERCENT OF FEMALES SINGLE ON SEX RATIO

| Region | Explained sum of squares and mean square variance | Errors of estimate | | | $P[F]$ |
		Unexplained sum of squares	Degrees of freedom	Mean square variance	
Northeast........	17.4	31.6	10	3.16	$P[F_{1.10} = 5.51] < .05$
Southeast........	5.2	41.8	6	6.97	$P[F_{1.6} = .75] > .05$
Middle States....	13.6	20.4	7	2.91	$P[F_{1.7} = 4.67] > .05$
The West........	21.3	9.7	7	1.39	$P[F_{1.7} = 15.32] < .01$

Source: Table 74.

gions are clearly not the proper units to use in studying the relationship between X and Y. The use of metropolitan areas within regions gives us more information on the relationship between X and Y. Had the positive between-region relationship been significant while the negative within-region relationship was also significant, then we would have had the conclusion that differences in sex ratio between regions do not affect the

Table 78. ANALYSIS OF VARIANCE FOR TESTING THE BETWEEN-REGION REGRESSION OF Y ON X

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Between-region means................	294	3		
Between-region regression about $\bar{Y}$......	27.5	1	27.5	
Region means around between-region regression........................	133.5	2	66.8	.41

$P[F_{1.2} = .41] > .05$

Source: Table 74.

percent of females single in the same way that differences in the sex ratio between metropolitan areas in the same region affect the percent of females single. Our next step would be an effort to explain why these relationships were different.

Comparison of regressions (Question 6). In the preceding paragraph we have discussed briefly the situation that would exist if we were to find the between-class and the average within-class regressions both significant but of opposite signs. It is possible, of course, that these two regressions be significantly different from each other without having opposite signs.

When one of the regressions is not significant, as in our case, it is not too meaningful to ask if the regressions are significantly different from each other, but we will illustrate the test, nevertheless. Table 79 shows this test. The sums of squares listed in Table 79 have already been identified,

Table 79. ANALYSIS OF ERRORS OF ESTIMATE FROM THREE REGRESSIONS

Source of variation	Errors of estimate		
	Sum of squares	Degrees of freedom	Mean square variance
Total....................	454.9	36	
a. Region means...............	180.7	2	90.4
b. Average within region.........	133.5	33	4.05
c. Remainder................	140.7	1	140.7

$$F_{2.33} = \frac{a}{b} = \frac{90.4}{4.05} = 22.3; \quad P[F_{2.33} = 22.3] < .001$$

$$F_{1.33} = \frac{c}{b} = \frac{140.7}{4.05} = 34.7; \quad P[F_{1.33} = 34.7] < .001$$

Source: Table 74.

with the exception of the last which is obtained by subtracting from the total the sum of the two sums immediately above it.

The first F formed supplies a test which confirms the test of Table 78, that there is no significant regression for the region means. (Table 78 showed that the explained sum of squares was not significant, and this first F test in Table 79 shows that the variance estimated from the sum of squares unexplained by the regression for the region means is significantly greater than the corresponding variance for the within-region regression.) The second F tests the significance of the difference between the average within-region regression and the regression of region means. Since the first F is significant and shows that there is no real trend in the relationship between Y and X between regions, it is rather pointless to test to see if this nonexistent trend is significantly different from the average within-

region trend. Nevertheless, the F has been found and is seen to be significant, which, if the first F had been insignificant, would mean that the average within-region regression coefficient is significantly different from the regression coefficient between means. Three regression lines based on these three regression coefficients, all intersecting at the point $(\bar{X}, \bar{Y})$ are shown in Figure 45.

Figure 45. Three Regressions of Percent of Females Single (Y) on Sex Ratio (X). (Source: Table 74.)

The analysis tables above include all the analyses usually made in this simplest case of an analysis of covariance problem. There are, however, other possible variations. For instance, the Middle States might be excluded and a new average within-region regression and associated correlation coefficient computed or even all the analyses made again on the new basis.

SUGGESTED READINGS

Dixon, Wilfrid J., and Massey, Jr., Frank J., *Introduction to Statistical Analysis* (New York: McGraw-Hill, 1951), Chap. 12.

Fisher, R. A., *Statistical Methods for Research Workers*, 10th ed. (London: Oliver and Boyd, 1948), pp. 270–284.

Lindquist, E. F., *Statistical Analysis in Educational Research* (Boston: Houghton, 1940), Chap. 6.

McNemar, Quinn, *Psychological Statistics* (New York: Wiley, 1949), Chap. 15.

Snedecor, George W., *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), Chaps. 12 and 13.

Taves, Marvin J., "The Application of Analysis of Covariance in Social Science Research," *American Sociological Review*, 15 (June 1950), pp. 373–381.

Multiple and Partial Correlation and Regression

Function of the methods of this chapter. The results of the analysis in the last chapter have illustrated the situation to which many inquiries into relationships lead—the realization that we must take into account other relevant factors. The process of changing our state of knowledge from one comprising only the description of an observed total association between two characteristics to one encompassing a more complex description of their association with other relevant characteristics taken into account is a further function of statistics—a function in the larger problem of establishing relationships with predictive validity from data on associations. If on the basis of analysis and description of a total association we predict levels of incidence of one characteristic from known levels of incidence of the associated characteristic, our predictions will be valid within the range of error determined by the degree of the observed association *only* if in the universe for which the predictions are made the distributions and interassociations of these two characteristics and of *all other* relevant factors are identically the same as in the universe from which our original association was observed. In a dynamic world "all relevant conditions" are never identically repeated. Therefore, we must learn to take into account the interassociations of as many relevant factors as possible in order to provide a more flexible basis for prediction (that is, to provide a technique for prediction which will be valid when the "other relevant factors" are somewhat altered). If the other relevant factors are characteristics whose measures form quantitative variables and if the two characteristics in whose association the primary interest lies are likewise quantitative, the methods of multiple and partial correlation and regression afford a battery of techniques for taking into account the other factors. The number of quantitative characteristics whose interassociations may be analyzed by the methods is theoretically unlimited, but the addition of each variable increases greatly the labor of the computation procedures.

BASIC CONCEPTS OF MULTIPLE AND PARTIAL CORRELATION AND REGRESSION

Notation for more than two variables. For defining the basic concepts of multiple and partial correlation and regression, we shall speak in terms of the simplest case, that of three variables where the form of all the associations, total and partial, is assumed to be linear. Whereas in total associations, or even in partial associations involving no more than two variables, we designated the variables with different letters, X and Y, in associations involving more than two quantitative variables we shall designate them all with the same letter and differentiate them by subscripts. Thus, in the simplest case of three variables we shall designate them as X_1, X_2, X_3. It does not really matter which subscript is allotted to which variable, but it is customary to let X_1 represent the "dependent" variable (the one to be predicted), if there is one.

Coefficient of partial correlation and regression. The coefficient of partial correlation, designated as $r_{12.3}$, measures the degree of association between X_1 and X_2, when X_3 is "allowed for" or statistically "held constant." The nature of the partial association between X_1 and X_2 is described by the partial regression coefficient, $b_{12.3}$, which indicates the average amount of change in level of incidence in X_1 associated with a unit change in X_2 when X_3 is held constant. Similarly the nature of the partial association between X_1 and X_3 is described by the partial regression coefficient, $b_{13.2}$, which indicates the average amount of change in level of incidence in X_1 associated with a unit change in X_3 when X_2 is held constant.

The multiple regression equation and the coefficient of multiple correlation. The multiple regression equation for estimating X_1 is constructed from the partial regression coefficients and one constant, $a_{1.23}$. For the case of three variables the multiple regression equation for predicting values of X_1 from given values of X_2 and X_3 is

$$X_{c1.23} = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \qquad (1)$$

The coefficient of multiple correlation, designated as $R_{1.23}$, measures the degree of association between X_1 and the other two variables considered simultaneously. The standard error of estimate of equation (1) is a function of $R_{1.23}$ in the same way that the standard error of estimate of a total regression equation,

$$X_{c1} = a_{12} + b_{12}X_2 \qquad (2)$$

is a function of the total correlation coefficient, r_{12}.

Differentiation of function of multiple and partial correlation and regression methods. When the emphasis of an inquiry lies in the net

association between two variables X_1 and X_2, then we can think of the third factor X_3 as merely a disturbing factor whose effects are to be eliminated. In such a case the coefficient of partial correlation, $r_{12.3}$, is used to measure the degree of association between X_1 and X_2 when X_3 is held constant. The partial coefficient, $r_{12.3}$, may be greater or smaller in absolute value than the corresponding total coefficient, r_{12}, according to whether the effect of X_3 (when not taken into account) obscures or exaggerates the net association between the two. Of course, if X_1 and X_2 are not associated with X_3 (that is, if r_{13} and r_{23} are both zero), the value of $r_{12.3}$ will be exactly the same as that for r_{12}. As a measure of the nature of the partial association (where primary interest lies in the association of X_1 and X_2), the partial regression coefficient, $b_{12.3}$, is used. We do not use $b_{12.3}$ to form a regression equation similar to (2), however, for estimating values of X_1 from given values of X_2. Such an equation would be applicable only in a universe where every unit had the same value of X_3 and, therefore, would rarely be practicable. For the partial association between X_1 and X_2 with X_3 held constant we can describe all four aspects of association—existence, direction, degree, and nature—but the description does not afford a basis of estimating X_1, given only information about X_2.

Where the emphasis of the inquiry lies in estimating one of the variables, X_1, from information on both of the other variables, X_2 and X_3— that is, where we are considering X_3 not simply a disturbing factor to be eliminated but useful additional information for prediction—the multiple regression equation (1) is our chief aim. It supplies a method for combining all the predictive value of the two total regression equations,

$$X_{c1.2} = a_{12} + b_{12}X_2$$
$$X_{c1.3} = a_{13} + b_{13}X_3$$

The regression coefficients used in the multiple regression equation must be partial and not total b's, for each must indicate the amount of change in X_1 associated with a unit change in the variable it modifies *when the disturbing effect of the other variable has been taken into account*. Thus, we see that the multiple regression equation utilizes the partial b's for estimating, although they could not be used for estimating before being combined into one equation.

The coefficient of multiple correlation, $R_{1.23}$, is the actual total r which would be obtained for the two series of values X_1 and $X_{c1.23}$. It measures the degree of association between the "dependent" variable and the other two combined. Just as the construction of the multiple regression equation is a way of combining two total regression equations, the coefficient of multiple correlation $R_{1.23}$ is a way of combining two total

coefficients of correlation r_{12} and r_{13}. In either case it is necessary to use corresponding partial coefficients in the process of combining. Therefore, the methods and procedures for multiple and partial correlation are quite intertwined.

Symmetry and asymmetry of coefficients. We should note which of these coefficients are symmetrical with respect to the variables involved, and which have different meanings for the "dependent" and "independent" variables. The partial r's are symmetrical with respect to the two variables whose subscripts appear before the period. For instance $r_{12.3}$ measures the degree of association between X_1 and X_2, with X_3 held constant, with just the same meaning for X_1 as for X_2, regardless of which is being considered the dependent variable. The coefficients of regression, however, always have reference to the predicting of one particular variable. For instance, $b_{12.3}$ is different in value and meaning from $b_{21.3}$ even though $r_{12.3}$ is identical with $r_{21.3}$; $b_{12.3}$ refers to the regression for predicting X_1 and $b_{21.3}$ refers to the regression for predicting X_2.

Coefficients to be computed in one problem. We shall list all the coefficients of correlation and regression possible in a three variable problem in multiple and partial correlation analysis. First, there are three total r's,

$$r_{12} \qquad r_{13} \qquad r_{23}$$

For the predicting of each variable from each of the others singly, there are two pairs of a's and b's, one pair for prediction from each of the other variables,

a_{12}	b_{12}		a_{21}	b_{21}	a_{31}	b_{31}
a_{13}	b_{13}		a_{23}	b_{23}	a_{32}	b_{32}

There are three partial r's,

$$r_{12.3} \qquad r_{13.2} \qquad r_{23.1}$$

(which could also be written as $r_{21.3}$, $r_{31.2}$, and $r_{32.1}$).

For the predicting of each variable from both of the others together, there is one set of regression coefficients,

$$a_{1.23} \quad b_{12.3} \quad b_{13.2} \qquad a_{2.13} \quad b_{21.3} \quad b_{23.1} \qquad a_{3.12} \quad b_{31.2} \quad b_{32.1}$$

And for each set of regression coefficients there is a multiple coefficient of correlation which measures their success at predicting,

$$R_{1.23} \qquad R_{2.13} \qquad R_{3.12}$$

(which could also be written $R_{1.32}$, $R_{2.31}$, $R_{3.21}$).

There are other related summarizing measures such as standard errors of estimates, coefficients of determination, etc., but the 15 total coefficients

and the 15 multiple and partial coefficients are the essential summarizing measures for describing the interassociations. If the problem is such that only one of the variables can meaningfully be considered a "dependent" variable, then the number of total coefficients needed is reduced to six or seven as is also the number of multiple and partial coefficients.

Choice of sequences of computation procedures. There are several choices of computation procedures for the multiple and partial coefficients. One may compute the partial r's directly from the total r's, and the partial b's directly from the total b's, and then compute the multiple R from the total and partial r's. Or one may compute the partial b's directly from the sums of the original measures, the sums of their squares, and the sums of their products; and then, by computing "explained" sums of squares from the b's and sums of products, find both partial and multiple coefficients of correlation. Or, as still another alternative, one may use combinations or variations of these two major sequences of procedure. In practice the choice of procedures depends upon which coefficients are desired for inquiry into a particular problem and upon what computations have already been made in investigating the total association. We shall illustrate the two major sequences for three variables. For analyzing the interassociations of a greater number of variables, the reader is referred to the suggested readings at the end of this chapter, in which extensions of the methods presented here will be found for any number of variables.

Order of coefficients. In problems with more than three variables it is necessary to distinquish the different orders of partial regression and correlation coefficients. The convention is to define the order of the partial coefficients as equal to the number of variables held constant, which is indicated in each partial regression or correlation coefficient by the number of subscripts to the right of the point. For instance, in a three variable problem, the partial coefficients, such $r_{12.3}$ and $b_{12.3}$ are called first order coefficients because one variable, X_3, as indicated in the subscript, is held constant. In a four variable problem there are partial coefficients of the second order, such as $r_{12.34}$ and $b_{12.34}$ where both X_3 and X_4 are held constant. If we apply the convention to total coefficients, such as r_{12} and b_{12}, where no variables are held constant, we see the reason that these total coefficients are often called zero order coefficients.

SEQUENCE OF COMPUTATION PROCEDURES UTILIZING COEFFICIENTS OF TOTAL CORRELATION AND REGRESSION

The problem. In Chapter 23 we examined the relationship between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators in the 31 economic areas of North Carolina, South Carolina, and Georgia. We

found a correlation coefficient of $-.515$ between these two variables, indicating that economic areas having large proportions of their farms operated by Negroes tend to have small percentages of farm homes with running water. A consideration of this situation suggests that these two variables, lack of running water in home and nonwhiteness, might be associated because of economic factors. At any rate, it would seem reasonable to ask the question, does a lower economic status of nonwhites explain this smaller percentage of homes with running water? We will attempt to answer this and other questions in the present chapter.

The data. For illustrating the procedures of multiple and partial correlation and regression analysis we shall use the data of Chapter 23 (listed in Table 55) and shall add the additional variable of median value of farm products sold or used by farm households for each of the 31 economic areas. Table 80 repeats the data of Chapter 23 and also gives the new variable. In this table though we have changed the designations. We now designate the percent of farms reporting running water in the dwelling as X_1 or sometimes by simply the numeral 1. The number of nonwhite farm operators per 100 white farm operators is designated as X_2, and the median value of farm products sold or used is designated as X_3. In the course of the problem we shall be interested in all the coefficients which do not "hold constant" the characteristic X_1, the percent of farms reporting running water in the dwelling. Therefore, in addition to the total coefficients, we shall want to compute the two partial r's,

$$r_{12.3} \qquad r_{13.2}$$

the two partial regression coefficients,

$$b_{12.3} \qquad b_{13.2}$$

the constant for the multiple regression equation,

$$a_{1.23}$$

and the coefficient of multiple correlation,

$$R_{1.23}$$

Computation of measures descriptive of the total associations. To emphasize the new procedures we shall not dwell upon the procedures for the total associations, but we shall assume that these procedures are familiar from Chapters 23 and 24. The following computations are obtainable from Table 80:

$$N = 31$$

$\Sigma X_1 = 452.3$	$\Sigma X_2 = 1{,}953$	$\Sigma X_3 = 412.5$
$\Sigma X_1^2 = 7{,}529.03$	$\Sigma X_2^2 = 209{,}699$	$\Sigma X_3^2 = 6{,}651.61$
$\Sigma X_1 X_2 = 23{,}868.7$	$\Sigma X_2 X_3 = 28{,}387.8$	$\Sigma X_1 X_3 = 5{,}381.01$

Table 80. PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING (X_1), NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X_2), AND MEDIAN VALUE OF FARM PRODUCTS SOLD OR USED PER FARM (X_3) IN ECONOMIC AREAS OF NORTH CAROLINA, SOUTH CAROLINA, AND GEORGIA, 1945

Economic area	Percent of farms reporting running water in dwelling	Nonwhite farm operators per 100 white farm operators	Median value of farm products sold or used in farm households [a] (hundreds of dollars)
North Carolina			
1 and A	28.0	1	6.9
2	16.8	4	5.6
3, B and C	17.5	27	19.8
4a	20.4	12	9.9
4b	27.4	13	9.3
5 and D	20.2	28	11.3
6 and E	9.7	46	23.4
7	5.8	125	24.1
8	7.9	72	29.2
9	9.0	106	22.1
10	8.5	39	18.4
11	7.8	48	18.3
South Carolina			
1	12.9	12	8.7
2	19.6	43	9.9
3	12.8	76	9.7
4	13.5	104	9.1
5 and A	17.3	62	11.3
6	10.6	180	13.9.
7	7.3	93	21.5
8 and B	9.7	176	5.4
Georgia			
1 and A	14.3	8	9.3
2	18.3	1	5.9
3 and B	18.6	12	8.6
4a	12.5	65	10.1
4b	13.5	123	8.8
5 and C	20.6	75	9.9
6	9.6	73	14.6
7a	14.4	195	15.5
7b	15.2	69	18.0
8	13.9	32	15.5
9 and D	18.7	33	8.5

[a] 1944

Source: Donald J. Bogue, *State Economic Areas: A description of the procedure used in making a functional grouping of the counties of the United States* (Washington: Government Printing Office, 1951), Table B.

From these data the intermediate and summarizing measures listed in Table 81 have been obtained. We shall use the total r's and b's of Table 81 to evaluate the formulas to be given in this first sequence of computations for the partial and multiple coefficients.

Computation of partial r's. The three possible partial r's may be obtained from the three total r's, but since we do not consider it meaningful to hold variable 1 constant, we shall compute only two of them. We want to know $r_{12.3}$, which measures the degree of association between the percent of farms reporting running water in the dwelling and the number of nonwhite farm operators per 100 white farm operators when median

Table 81. INTERMEDIATE AND SUMMARIZING MEASURES FOR FOUR REGRESSIONS FROM DATA OF TABLE 80

	Regression of X_1 on X_2	Regression of X_1 on X_3	Regression of X_2 on X_3	Regression of X_3 on X_2
	$\Sigma x_1^2 = 929.83$	$\Sigma x_1^2 = 929.83$	$\Sigma x_2^2 = 86,660$	$\Sigma x_3^2 = 1,162.70$
	$\Sigma x_2^2 = 86,660$	$\Sigma x_3^2 = 1,162.70$	$\Sigma x_3^2 = 1,162.70$	$\Sigma x_2^2 = 86,660$
	$\Sigma x_1 x_2 = -4,626.20$	$\Sigma x_1 x_3 = -637.498$	$\Sigma x_2 x_3 = 2,400.30$	$\Sigma x_3 x_2 = 2,400.30$
$r_{12} = \dfrac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \Sigma x_2^2}} = -.5154$		$r_{13} = -.6131$	$r_{23} = .2391$	$r_{32} = .2391$
$b_{12} = \dfrac{\Sigma x_1 x_2}{\Sigma x_2^2} = -.05338$		$b_{13} = -.5483$	$b_{23} = 2.0644$	$b_{32} = .02770$
$a_{12} = 17.95$		$a_{13} = 21.89$	$a_{23} = 35.53$	$a_{32} = 11.56$
$X_{c1.2} = 17.95 - .05338 X_2$		$X_{c1.3} = 21.89 - .5483 X_3$	$X_{c2.3} = 35.53 + 2.064 X_3$	$X_{c3.2} = 11.56 + .0277 X_2$

Source: Table 80.

value of farm products is held constant and also $r_{13.2}$, which measures the degree of association between the percent of farms reporting running water in the dwelling and the median value of farm products when the number of nonwhite farm operators per 100 white farm operators is held constant. Both of these may be obtained by substituting values of the total r's in the formula

$$r_{1j.k} = \frac{r_{1j} - r_{1k} r_{jk}}{\sqrt{(1 - r_{1k}^2)(1 - r_{jk}^2)}} \tag{3}$$

There are various tables listed at the end of this chapter which facilitate the evaluation of (3). If they are not available, the evaluation can be quickly performed on the calculator or by means of logarithms. In any case, the subscript "1" refers to the "dependent" variable, the subscript "j" to the other variable whose net association with "1" is being investigated, and the subscript "k" to the variable which is being held constant. To find a measure of the degree of association between the percent of farm homes reporting running water (1) and the number of nonwhite farm operators per 100 white farm operators (2) when median

value of farm products (3) is held constant, we evaluate formula (3) with data of Table 81 as follows,

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}}$$

$$= \frac{-.5154 - (-.6131)(.2391)}{\sqrt{[1 - (-.6131)^2][1 - (.2391)^2]}}$$

$$= -.4808$$

From comparing the absolute size of $r_{12.3} = -.4808$ with the absolute size of the total r, $r_{12} = -.5154$, we find that the effect of holding constant variable 3, with which variable 1 is negatively associated and variable 2 is positively associated, decreases the degree of association between 1 and 2. In terms of our problem this means that some of the apparent association between the percent of farm homes with running water and the number of nonwhite farm operators per 100 white farm operators is accounted for or "explained" by differences in the value of farm products.

Next we examine the net association between the percent of farm homes having running water and the value of farm products, holding constant the number of nonwhite farm operators per 100 white farm operators. Again we evaluate formula (3) from the data of Table 81, this time considering variable 3 as j and variable 2 as k, thus,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r^2_{12})(1 - r^2_{23})}}$$

$$= \frac{-.6131 - (-.5154)(.2391)}{\sqrt{[1 - (-.5154)^2][1 - (.2391)^2]}}$$

$$= -.5888$$

As before, the partial correlation between the percent of farms with running water and one of the other variables $(-.5888)$ is smaller in absolute value than the total correlation $(-.6131)$ between the two.

Computation of the coefficient of multiple correlation. Let us develop a formula for the coefficient of multiple correlation between variable 1 and variables 2 and 3 considered simultaneously. As with total correlation, the square of the coefficient of multiple correlation, $R^2_{1.23}$, is the proportion of the variation in variable 1 explained by variables 2 and 3. At first thought one might try to obtain $R^2_{1.23}$ by adding the squared total coefficients, r^2_{12} and r^2_{13}, since proportions of variation explained are additive. This will not work, however, since we have seen that there is overlapping in the explaining of variation by the two variables. What we should add is the square of one of the total coefficients, and then an expression representing the additional amount the other variable explains.

Actually the formula for $R^2_{1.23}$ is

$$R^2_{1.23} = r^2_{12} + r^2_{13.2}(1 - r^2_{12}) \tag{4}$$

The first term on the right represents the proportion of the total variation in variable 1 explained by variable 2. The second term represents the proportion of the remaining variation in variable 1 explained by variable 3 after variable 2 had explained all it could. The second term is a product of two proportions. The proportion $1 - r^2_{12}$ represents the proportion of the total variation in 1 that was left unexplained by variable 2. The proportion $r^2_{13.2}$ is the proportion of $1 - r^2_{12}$ which variable 3 explains. In fact the coefficient of partial association, $r_{13.2}$, may be defined as the square root of the proportion of variation in variable 1 left unexplained by variable 2 which variable 3 explains.

Substituting in (4) the r's of our example, we have

$$R^2_{1.23} = (-.5154)^2 + (-.5888)^2[1 - (-.5154)^2]$$
$$= .5202$$
$$R_{1.23} = .7212$$

The coefficient of multiple correlation does not have any direction since the concept of direction is not applicable to association between one variable and several others considered simultaneously. The coefficient of multiple correlation often is a combined measure of positive and negative relationships. The multiple correlation coefficient cannot be smaller in absolute value than any of the coefficients of total correlation between the independent variable and the others. This is necessarily so because $R^2_{1.23}$ represents the proportion of variation in variable 1 explained by variables 2 and 3, and this proportion cannot be smaller than the proportion explained by either variable alone, r^2_{12} or r^2_{13}.

Formula (4) may be written in an alternate form, thus,

$$R^2_{1.23} = r^2_{13} + r^2_{12.3}(1 - r^2_{13}) \tag{5}$$

indicating that the sequence in which we consider the two variables, 2 and 3, is reversed; that is, we first let variable 3 explain all of the variation of 1 it can and then let 2 explain the proportion, $r^2_{12.3}$ of what is left. Substitution of the data of our example in (5) gives

$$R^2_{1.23} = (-.6131)^2 + (-.4808)^2[1 - (-.6131)^2]$$
$$= .5202$$
$$R_{1.23} = .7212$$

which affords a check on previous computations.

Extension of procedures for four variables. We have seen how to compute the coefficients of partial correlation from coefficients of total

correlation—that is, we have computed coefficients of the first order from coefficients of zero order. An extension of formula (3) makes it possible in problems of more than three variables to compute partial coefficients of any order from the partial coefficients of lower orders. For instance, if there were a fourth variable in this problem, $r_{12.34}$, a partial coefficient of the second order could be computed from partial coefficients of the zero and first orders by the relation,

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \qquad (6)$$

Similarly, the formula for $R_{1.234}^2$ becomes

$$R_{1.234}^2 = r_{12}^2 + r_{13.2}^2(1 - r_{12}^2) + r_{14.23}^2[1 - r_{12}^2 - r_{13.2}^2(1 - r_{12}^2)] \qquad (7)$$

where the third term on the right represents the proportion of variation in 1 unexplained by 2 and 3 which 4 explains.

Computation of partial regression coefficients. The first order partial regression coefficients may be obtained from the zero order regression coefficients just as the first order partial correlation coefficients are obtained from the zero order correlation coefficients. To construct a multiple regression equation for predicting values of X_1 from known values of X_2 and X_3, we shall need to know two first order partial regression coefficients, $b_{12.3}$ and $b_{13.2}$. These may be obtained from the formula,

$$b_{1j.k} = \frac{b_{1j} - b_{1k}b_{kj}}{1 - b_{jk}b_{kj}} \qquad (8)$$

To obtain the partial regression coefficient of the percent of farm homes with running water on the number of nonwhite farm operators per 100 white farm operators with the median value of farm products held constant, we substitute data of Table 81 in formula (8), thus,

$$b_{12.3} = \frac{b_{12} - b_{13}b_{32}}{1 - b_{23}b_{32}}$$

$$= \frac{-.05338 - (-.5483)(.02770)}{1 - (2.0644)(.02770)}$$

$$= -.04051$$

And to obtain the partial regression coefficient of the percent of farm homes with running water on the median value of farm produce when the number of nonwhite farm operators per 100 white farm operators is held constant, we substitute data of Table 81 in the same formula, thus,

$$b_{13.2} = \frac{b_{13} - b_{12}b_{23}}{1 - b_{32}b_{23}}$$

$$= \frac{-.5483 - (-.05338)(2.0644)}{1 - (2.0644)(.02770)}$$

$$= -.46468$$

Computation of the constant term in the multiple regression equation. The constant term of the multiple regression equation for predicting values of X_1 from values of X_2 and X_3 is designated as $a_{1.23}$. This term is a function of the values of the means of the three distributions and of the two partial regression coefficients. Its formula in terms of means is

$$a_{1.23} = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3 \tag{9}$$

or in terms of sums of measures it is

$$a_{1.23} = \frac{\Sigma X_1 - b_{12\,3}\Sigma X_2 - b_{13.2}\Sigma X_3}{N} \tag{10}$$

Substitution of the partial regression coefficients just determined and of data listed on page 504 gives

$$a_{1.23} = \frac{452.3 - (-.04051)(1953) - (-46468)(412.5)}{31}$$

$$= 23.33$$

Construction of the multiple regression equation. Finally, substitution of this value and of the values for the partial regression coefficients in equation (1) gives us as the multiple regression equation for estimating values of X_1 from known values of X_2 and X_3,

$$X_{c1.23} = 23.33 - .04051X_2 - .46468X_3 \tag{11}$$

SEQUENCE OF COMPUTATION PROCEDURES NOT REQUIRING PREVIOUS COMPUTATION OF TOTAL COEFFICIENTS

Before inquiring into the significance of the multiple and partial coefficients, we shall present an alternate sequence of computations which will lead to slightly more accurate determination of the coefficients in equation (11). In this sequence we shall compute $a_{1.23}$, $b_{12.3}$, and $b_{13.2}$ directly from the gross sums of measures, of their squares, and of their products (ΣX_1, ΣX_1^2, $\Sigma X_1 X_2$ etc.) instead of from coefficients of lower order, thus preventing exaggeration of errors of rounding.

The Doolittle method. This second sequence of operations requires the simultaneous solution of as many equations in as many unknowns as

there are variables—in our case the solution of three simultaneous equations with $a_{1.23}$, $b_{12.3}$, and $b_{13.2}$ as the unknowns to be solved for. Any of the algebraic methods of solving three simultaneous equations are satisfactory —elimination of unknowns by alternate procedures or solution by determinants. The number of simultaneous equations to be solved may be reduced by one if the equations are set up in terms of deviations from means rather than in terms of original measures.

Certain features of the equations to be solved make possible a systematic arrangement of procedures known as the "Abbreviated Doolittle Method." This method is especially satisfactory to follow if there are more than three variables. We will show the general form of the computations for four variables, and from this form the student should be able to make extensions for any number of variables. We will illustrate the computations for our three variable problem.

The normal equations. In a four variable problem the regression coefficients we wish to find are $a_{1.234}$, $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$. The equations which we solve to find these coefficients are called the normal equations. They are as follows:

$$Na_{1.234} + \Sigma X_2 b_{12.34} + \Sigma X_3 b_{13.24} + \Sigma X_4 b_{14.23} = \Sigma X_1 \quad (12)$$
$$\Sigma X_2 a_{1.234} + \Sigma X^2_2 b_{12.34} + \Sigma X_2 X_3 b_{13.24} + \Sigma X_2 X_4 b_{14.23} = \Sigma X_1 X_2 \quad (13)$$
$$\Sigma X_3 a_{1.234} + \Sigma X_2 X_3 b_{12.34} + \Sigma X^2_3 b_{13.24} + \Sigma X_3 X_4 b_{14.23} = X_1 X_3 \quad (14)$$
$$\Sigma X_4 a_{1.234} + \Sigma X_2 X_4 b_{12.34} + \Sigma X_3 X_4 b_{13.24} + \Sigma X^2_4 b_{14.23} = \Sigma X_1 X_4 \quad (15)$$

Solution of normal equations by the abbreviated Doolittle method. The whole equation (12) and the parts of equations (13), (14), and (15) that are needed in the solution by the Doolittle method are shown in the first four lines of Table 82. The remainder of Table 82, with the exception of column (6), shows how to make the necessary computations in order to obtain the regression equation. Column (6) serves as a check on the computations. The entries of column (6) are achieved as shown on Table 82. These entries serve as a check because each one must equal the sum of all the other entries in the same row with it except for rounding errors. For example,

$$j_7 = C_2 + C_3 + C_4 + C_5$$

After using Table 82 and seeing the pattern of the computations, a student should be able to go through an abbreviated Doolittle solution with a larger number of variables. Each additional variable adds one column and two rows to the computations.

It is possible to expand the Doolittle method to include a "back" solution, which involves the computation of the higher order partial

Table 82. GENERAL FORM FOR OBTAINING THE COEFFICIENTS OF THE MULTIPLE REGRESSION EQUATION BY THE ABBREVIATED DOOLITTLE METHOD FOR SOLVING THE NORMAL EQUATIONS

Rows	$a_{1.234}$ (1)	$b_{12.34}$ (2)	$b_{13.24}$ (3)	$b_{14.23}$ (4)	Constant (5)	Check (6)
1.	N	ΣX_2	ΣX_3	ΣX_4	ΣX_1	j_1 = sum of row entries
2.		ΣX_2^2	$\Sigma X_2 X_3$	$\Sigma X_2 X_4$	$\Sigma X_1 X_2$	j_2 = sum of row entries plus omitted entries
3.			ΣX_3^2	$\Sigma X_3 X_4$	$\Sigma X_1 X_3$	j_3 = sum of row entries plus omitted entries
4.				ΣX_4^2	$\Sigma X_1 X_4$	j_4 = sum of row entries plus omitted entries
5.	$A_1 = N$	$A_2 = \Sigma X_2$	$A_3 = \Sigma X_3$	$A_4 = \Sigma X_4$	$A_5 = \Sigma X_1$	$j_5 = j_1$
6.	$B_1 = 1$	$B_2 = \dfrac{A_2}{A_1}$	$B_3 = \dfrac{A_3}{A_1}$	$B_4 = \dfrac{A_4}{A_1}$	$B_5 = \dfrac{A_5}{A_1}$	$j_6 = \dfrac{j_5}{A_1}$
7.		$C_2 = \Sigma X_2^2 - A_2 B_2$	$C_3 = \Sigma X_2 X_3 - A_3 B_2$	$C_4 = \Sigma X_2 X_4 - A_4 B_2$	$C_5 = \Sigma X_1 X_2 - A_5 B_2$	$j_7 = j_2 - j_5 B_2$
8.		$D_2 = 1$	$D_3 = \dfrac{C_3}{C_2}$	$D_4 = \dfrac{C_4}{C_2}$	$D_5 = \dfrac{C_5}{C_2}$	$j_8 = \dfrac{j_7}{C_2}$
9.			$E_3 = \Sigma X_3^2 - A_3 B_3 - C_3 D_3$	$E_4 = \Sigma X_3 X_4 - A_4 B_3 - C_4 D_3$	$E_5 = \Sigma X_1 X_3 - A_5 B_3 - C_5 D_3$	$j_9 = j_3 - j_5 B_3 - j_7 D_3$
10.			$F_3 = 1$	$F_4 = \dfrac{E_4}{E_3}$	$F_5 = \dfrac{E_5}{E_3}$	$j_{10} = \dfrac{j_9}{E_3}$
11.				$G_4 = \Sigma X_4^2 - A_4 B_4 - C_4 D_4 - E_4 F_4$	$G_5 = \Sigma X_1 X_4 - A_5 B_4 - C_5 D_4 - E_5 F_4$	$j_{11} = j_4 - j_5 B_4 - j_7 D_4 - j_9 F_4$
12.				$H_4 = 1$	$H_5 = \dfrac{G_5}{G_4}$	$j_{12} = \dfrac{j_{11}}{G_4}$

$b_{14.23} = H_5$

$b_{13.24} = -b_{14.23}F_4 + F_5$

$b_{12.34} = -b_{13.24}D_3 - b_{14.23}D_4 + D_5$

$a_{1.234} = -b_{12.34}B_2 - b_{13.24}B_3 - b_{14.23}B_4 + B_5$

$X_{c1.234} = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$

correlations. However, it is felt that the method presented here is shorter unless one also desires the other products of the "back" solution.[1]

The organization of Table 82 assumes that the computer has access to a modern calculating machine that will add and substract products automatically. If such a machine is not available, it will be necessary to write down the separate products involved and then add them.

Illustration of computations of Abbreviated Doolittle Solution. Table 83 shows the computations for obtaining regression equation (11) directly

Table 83. COMPUTATIONS FOR OBTAINING THE MULTIPLE REGRESSION EQUATION BY THE ABBREVIATED DOOLITTLE METHOD FOR SOLVING THE NORMAL EQUATIONS

Row	$a_{1.23}$	$b_{12.3}$	$b_{13.2}$	Constant	Check
1	31	1953	412.5	452.3	2,848.8
2		209,699	28,387.8	23,868.7	263,908.5
3			6,651.61	5,381.01	40,832.92
4	31	1953	412.5	452.3	2,848.8
5	1	63	13.306452	14.590323	91.89677
6		86,660	2400.3	$-4,626.2$	84,434.1
7		1	.02769789	$-.05338333$	.974314
8			1,096.2153	-509.36226	586.8531
9			1	$-.464655$	.535344

$b_{13.2} = -.464655$

$b_{12.3} = -(-.464655)(.02769798) - .0533833 = -.040513$

$a_{1.23} = -(-.040513)(63) - (-.464655)(13.306452) - (-14.590323) = 23.3256$

$X_{c1.23} = 23.3256 - .040513X_2 - .464655X_3$

Source: Table 82 and data from page 504.

from the original sums shown on page 504. It is seen that the regression coefficients computed by the abbreviated Doolittle solution agree with those previously computed to four significant digits.

Comparison of procedures. In general the abbreviated Doolittle solution is the most direct way of obtaining the multiple regression solution because original sums are used rather than intermediate computed measures. Either method can be made to yield accuracy to any required degree by carrying the intermediate divisions to a sufficient number of decimal places. In a three variable problem the choice of

---

[1] See Harold Hotelling, "Some New Methods in Matrix Calculation," *The Annals of Mathematical Statistics,* 14 (1943), pp. 1-12; and Harold Hotelling, "Further Points on Matrix Calculation and Simultaneous Equations," 14 (1943), p. 440.

sequence should be determined by what one wishes to obtain and what preliminary computations have already been made. If nothing is desired except the multiple regresssion equation, one should use the second sequence and obtain it directly. If all the total coefficients have been computed in preliminary investigations, the first sequence is shorter. Suppose, however, that the second sequence has been used, and one later decides he needs some of the other measures. The following procedures explain how they can be obtained from the partial regression coefficients found above. These procedures do, however, require the computation of the following sums of squared deviations and product deviations, shown in Table 81:

$$\Sigma x_1^2 = 929.83 \qquad\qquad \Sigma x_2^2 = 86,660$$
$$\Sigma x_1 x_2 = -4,626.20 \qquad\qquad \Sigma x_3^2 = 1,162.70$$
$$\Sigma x_1 x_3 = -637.498$$

Computation of explained sums of squares. It will be remembered that the expression for the explained sum of squares in a total regression, which we shall now designate as $\Sigma x_{c1.2}^2$, is equal to b_{12} times the sum of the product deviations, thus,

$$\textit{Sum of squares in 1}$$
$$\textit{explained by total re-} = \Sigma x_{c1.2}^2 = b_{12}\Sigma x_1 x_2 \qquad (16)$$
$$\textit{gression of 1 on 2}$$

Similarly in a multiple regression the explained sum of squares is given by the formula,

$$\textit{Sum of squares in 1}$$
$$\textit{explained by multiple} = \Sigma x_{c1.23}^2 = b_{12.3}\Sigma x_1 x_2 + b_{13.2}\Sigma x_1 x_3 \qquad (17)$$
$$\textit{regression of 1 on 2 and 3}$$

Substitution in (17) of the partial b's (rounded) and the sums of the product deviations gives

$$\Sigma x_{c1.23}^2 = (-.040513)(-4,626.2) + (-.46466)(-637.498)$$
$$= 483.641$$

Computation of the coefficient of multiple correlation. The square of the coefficient of multiple correlation is equal to the proportion of total variation in X_1 explained by its multiple regression on X_2 and X_3, that is,

$$R_{1.23}^2 = \frac{\Sigma x_{c1.23}^2}{\Sigma x_1^2} \qquad (18)$$

Substituting our data gives

$$R_{1.23}^2 = \frac{483.641}{929.829} = .52014$$

$$R_{1.23} = .7212$$

Computation of the standard error of estimate of the multiple regression equation. As in total regression the sum of squares unexplained by a multiple regression equation is equal to the total sum of squares minus the explained sum of squares. Designating the unexplained sum of squares as $\Sigma x_{s1.23}^2$, we have

$$\Sigma x_{s1.23}^2 = \Sigma x_1^2 - \Sigma x_{c1.23}^2 \qquad (19)$$

The estimate of the square of the standard error of estimate is obtained by dividing this unexplained sum of squares by $N - m$, where m is the number of constants in the regression equation, three in our case, thus,

$$\hat{\sigma}_{s1.23} = \sqrt{\frac{\Sigma x_{s1.23}^2}{N - m}} \qquad (20)$$

Substitution in (19) and (20) gives

$$\hat{\sigma}_{s1.23} = \sqrt{\frac{929.829 - 483.641}{31 - 3}} = \sqrt{15.9353} = 3.992$$

Computation of sum of squares explained by two total regressions. In order to obtain the partial r^2's as proportions of variation explained, we must know the sum of squares explained in the two total regressions, X_1 on X_2 and X_1 on X_3. These may be obtained directly from the usual formula (16) for the explained sum of squares if we know the total b's. Since the b's have not been computed, however, we use the equivalent formula,

$$\Sigma x_{c1.2}^2 = \frac{(\Sigma x_1 x_2)^2}{\Sigma x_2^2} \qquad (21)$$

Substitution in (21) gives as the sum of squares in 1 explained by total regression on 2,

$$\Sigma x_{c1.2}^2 = \frac{(-4{,}626.2)^2}{86{,}660} = 246.96$$

Modification of the subscripts of formula (21) to obtain the sum of squares in 1 explained by total regression on 3, gives

$$\Sigma x_{c1.3}^2 = \frac{(\Sigma x_1 x_3)^2}{\Sigma x_3^2} \qquad (22)$$

and substitution of our data in (22) gives

$$\Sigma x_{c1.3}^2 = \frac{(-637.498)^2}{1,162.70} = 349.53$$

Computation of partial r's. We can now compute the partial r's from the relations expressed in the definition of the square of the coefficient of partial correlation; that is, the square of the coefficient of partial correlation between X_1 and X_2 with X_3 held constant is the proportion of the variation in X_1 unexplained by X_3 which X_2 explains. Let us develop expressions for both the numerator and the denominator of this proportion. For the denominator we require the amount of variation in X_1 unexplained by X_3. This is, of course, equal to the total variation in X_1 minus the variation explained by the total regression on X_3. Since it describes the scatter of X_1 around the regression on X_3, we shall designate this unexplained variation as $\Sigma x_{1.3}$. Its computation is indicated by the formula,

$$\Sigma x_{s1.3}^2 = \Sigma x_1^2 - \Sigma x_{c1.3}^2 \tag{23}$$

which we recognize as the familiar analysis of variance relation that the unexplained sum of squares is equal to the total sum of squares minus the explained sum of squares.

The numerator of the proportion equal to $r_{12.3}^2$ must be the amount of variation in X_1 which X_2 explains *after* X_3 has explained all it can. We shall not compute this amount directly, but by subtraction. We know that $\Sigma x_{c1.23}^2$ is the total amount of variation in X_1 explained by the multiple regression on X_2 and X_3; we know that $\Sigma x_{c1.3}^2$ is the variation in X_1 explained by the total regression on X_3. Therefore, the excess of $\Sigma x_{c1.23}^2$ over $\Sigma x_{c1.3}^2$ must be the additional amount X_2 explains after X_3 has explained all it can, and our desired numerator is

Variation in 1 explained
by 2 in addition to that $= \Sigma x_{c1.23}^2 - \Sigma x_{c1.3}^2$
explained by 3

The required formula for $r_{12.3}^2$ can be made now from the numerator and denominator just developed,

$$r_{12.3}^2 = \frac{\Sigma x_{c1.23}^2 - \Sigma x_{c1.3}^2}{\Sigma x_1^2 - \Sigma x_{c1.3}^2} \tag{24}$$

Or if we first let 2 explain all it can and then express the proportion of unexplained which 3 explains, we obtain a similar formula for $r_{13.2}^2$,

$$r_{13.2}^2 = \frac{\Sigma x_{c1.23}^2 - \Sigma x_{c1.2}^2}{\Sigma x_1^2 - \Sigma x_{c1.2}^2} \tag{25}$$

Substitution of our data on "explained" and total sums of squares in (24) and (25) gives

$$r_{12.3}^2 = \frac{483.641 - 349.53}{929.829 - 349.53} = .2311$$

$$r_{12.3} = -.4807$$

$$r_{13.2}^2 = \frac{483.641 - 246.96}{929.829 - 246.96} = .3466$$

$$r_{13.2} = -.5887$$

The signs of the partial r's cannot be determined by these procedures, and we, therefore, give them the signs of their corresponding partial b's, which also measure direction of association.

Summary of second sequence of computation procedures. By this second sequence of procedures we have obtained the two partial regression coefficients, $b_{12.3}$ and $b_{13.2}$, the constant term of the multiple regression equation $a_{1.23}$, the coefficient of multiple correlation, $R_{1.23}$, and the two partial coefficients of correlations, $r_{12.3}$ and $r_{13.2}$, without using any of the coefficients of the total associations. It is true that in the computation of $R_{1.23}$, $r_{12.3}$, and $r_{13.2}$ we have used the intermediate sums of squared deviations and product deviations, Σx^2, Σx_2^2, Σx_3^2, $\Sigma x_1 x_2$, and $\Sigma x_1 x_3$; but if we wished $R_{1.23}$, $r_{12.3}$ and $r_{13.2}$ to be computed directly from gross sums of squares and products of the original measures, all that would be necessary would be to substitute the right side of the formulas for the sums of squared deviations and product deviations, such as,

$$\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{N}$$

$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - \frac{(\Sigma X_1)(\Sigma X_2)}{N}$$

and so on

in formulas (17), (21), and (22) and the resulting quantities in formulas (18), (24), and (25).

TESTS OF SIGNIFICANCE OF THE MULTIPLE AND PARTIAL CORRELATION AND REGRESSION COEFFICIENTS

As in the case of coefficients of total correlation and regression, we have a number of choices in testing the hypothesis that there is no correlation in the universe. And as before, a test of significance of a coefficient of partial correlation is also a test of the significance of its corresponding co-

efficient of partial regression. Similarly, a test of the coefficient of multiple correlation is a test of the significance of the improvement of estimated values made by using the multiple regression equation (and information on levels of X_2 and X_3) over estimated values, made by using as estimate simply the mean of the distribution of X_1.

Test of significance of coefficient of multiple correlation. We shall use analysis of variance tests for the significance of the coefficients and then show how to set up confidence limits for the multiple and partial r's using the Z transformation. Table 84 shows the test of significance of the coefficient of multiple correlation, $R_{1.23}$, which proves to be significant

Table 84. ANALYSIS OF VARIANCE IN PERCENT OF FARMS REPORTING RUN-NING WATER IN DWELLING (X_1) FOR TESTING THE SIGNIFICANCE OF THE MULTIPLE REGRESSION ON NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X_2) AND MEDIAN VALUE OF FARM PRODUCTS PER FARM (X_3) IN 31 ECONOMIC AREAS, 1945

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total...................	929.829	30		
Explained by multiple regression on 2 and 3.................	483.641	2	241.82	
				15.17
Unexplained by 2 and 3.........	446.188	28	15.94	

$P[F_{2.28} = 15.17] < .001$

Source: Table 80.

beyond the .001 level. Notice that the sum of squares explained by 2 and 3 (computed on page 514) is associated with two degrees of freedom representing the two partial regression coefficients in the multiple regression equation.

Tests of significance of partial correlation and regression coefficients. Table 85 shows the tests of significance of the coefficients of correlation, r_{13} and $r_{12.3}$. Both are significant, although the latter at a lower level. The sum of squares "unexplained by 3" in Table 85 is obtained by the usual subtraction of the "explained by 3" from the total. The "unexplained by 2 and 3" was computed in Table 84. The other sums of squares have all been computed in the immediately preceding pages, except the "additional explained by 2," which is found by subtracting the "unexplained by 2 and 3" from the "unexplained by 3" thus,

$$580.299 - 446.188 = 134.111$$

Note that different denominators are used for the two F's. The mean square variance 134.11 must be related to the variance estimated from the portion of 580.299 which 3 left unexplained. One should not subtract 134.111 from the total variation in X_1 for obtaining a sum of squares called "unexplained by the partial regression on X_2," because X_2 was not given a chance to explain any variation except that left unexplained by X_3, 580.299.

Table 85. ANALYSIS OF VARIANCE IN PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING (X_1) FOR TESTING THE SIGNIFICANCE OF THE TOTAL REGRESSION ON MEDIAN VALUE OF FARM PRODUCTS PER FARM (X_3) AND FOR TESTING THE SIGNIFICANCE OF THE PARTIAL REGRESSION ON NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X_2), 31 ECONOMIC AREAS, 1945

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total........................	929.829	30		
a. Explained by 3.............	349.53	1	349.53	
				17.47
b. Unexplained by 3..........	580.299	29	20.01	
c. Additional explained by 2....	134.111	1	134.11	
				8.41
d. Unexplained by 2 and 3.....	446.188	28	15.94	

$F_{1.29} = \dfrac{a}{b} = 17.47; \quad P[F_{1.29} = 17.47] < .001$

$F_{1.28} = \dfrac{c}{d} = 8.41; \quad P[F_{1.28} = 8.41] < .01$

Source: Table 80.

Table 86 shows similar tests for r_{12} and $r_{13.2}$. Note that the test for the partial correlation of X_1 with either of the other variables must be made in connection with the test for the total correlation of X_1 with the other variable.

Computation of confidence limits of the multiple and partial correlation coefficients. Confidence limits for the multiple and partial coefficients of correlation can be set up by the Z transformation through procedures the same as those for total coefficients with one exception. For partial r's the number of "effective" observations is decreased one by every variable held constant. Therefore, the standard error of Z given by the formula,

$$\sigma_Z = \sqrt{\frac{1}{N-3}} \tag{26}$$

is not applicable. In fact (26) is a special case of a more general formula,

$$\sigma_Z = \sqrt{\frac{1}{N - m - 1}} \qquad (27)$$

where N = number of cases
m = constants in the regression equation or degrees of freedom sacrificed in determining a coefficient of correlation

Table 86. ANALYSIS OF VARIANCE IN PERCENT OF FARMS REPORTING RUNNING WATER IN DWELLING (X_1) FOR TESTING THE SIGNIFICANCE OF THE TOTAL REGRESSION ON NUMBER OF NONWHITE FARM OPERATORS PER 100 WHITE FARM OPERATORS (X_2) AND FOR TESTING THE SIGNIFICANCE OF THE PARTIAL REGRESSION ON MEDIAN VALUE OF FARM PRODUCTS PER FARM (X_3), 31 ECONOMIC AREAS, 1945

Source of variation	Sum of squares	Degrees of freedom	Mean square variance	F
Total........................	929.829	30		
a. Explained by 2.............	246.96	1	246.96	
b. Unexplained by 2..........	682.869	29	23.55	10.49
c. Additional explained by 3....	236.681	1	236.68	
d. Unexplained by 2 and 3.....	446.188	28	15.94	14.85

$F_{1.29} = \dfrac{a}{b} = 10.49; \quad P[F_{1.29} = 10.49] < .01$

$F_{1.28} = \dfrac{c}{d} = 14.85; \quad P[F_{1.28} = 14.85] < .001$

Source: Table 80.

For $R_{1.23}$, $r_{12.3}$, and $r_{13.2}$, substitution in (27) gives

$$\sigma_Z = \sqrt{\frac{1}{31 - 3 - 1}} = \sqrt{\frac{1}{27}} = .1924501$$

The corresponding Z's for the three r's are found from Appendix Table G to be as follows,

r	Z
$R_{1.23} = .721$	.909725
$r_{12.3} = -.481$	.524285
$r_{13.2} = -.589$	.676134

The distance from each Z to its upper and lower 95-percent confidence limit is

$$1.96\sigma_z = 1.96(.1924501) = .377202$$

Subtracting and adding .37720 to each value of Z we obtain the confidence limits,

Z	Lower confidence limit	Upper confidence limit
.909725	.532523	1.286927
.524285	.147083	.901487
.676134	.298932	1.053336

Transforming these Z values to corresponding r's through inverse use of Appendix Table G, we have

r		Lower confidence limit	Upper confidence limit
$R_{1.23} =$	.721	.487	.858
$r_{12.3} =$	$-.481$	$-.146$	$-.717$
$r_{13.2} =$	$-.589$	$-.290$	$-.783$

These confidence limits are to be interpreted just as in the case of confidence limits of total r's. They imply the same type of tests of hypotheses about the universe from which the 31 observations may be considered a random sample.

The Z transformation may also be used for testing hypotheses of types IV and V, explained in Chapter 23, for multiple and partial correlation coefficients with modified formula (27) used for the standard error of Z. Again, however, it must be remembered that tests of type V, for determining the significance of the difference between two coefficients of correlation are based on the assumption that the coefficients have been observed in two independent samples. Such tests are not applicable to testing the significance of the difference between two partial r's of the same sample, such as $r_{12.3}$ and $r_{13.2}$.

Omission of summaries of descriptions of associations. We shall not summarize the results of the descriptions of the interassociations between X_1, X_2, and X_3, developed in this chapter, although such summaries could be made. Nor shall we differentiate between the description for the sample and the estimated description for the universe of possibilities, trusting by now that the student can do this for himself, given the coefficients, tests of significance, and confidence limits computed in this chapter. Such summaries and differentiations are not customarily presented in the reports of investigations of associations. We have used them heretofore not as examples to be followed in actual research reporting but for clarifying the meaning of the different aspects of association, their summarizing measures, and their tests of significance. As soon as they are fully understood by the student, they can be omitted.

This chapter should be considered merely an introduction to multiple and partial correlation and regression. In it we have taken up only the procedures for analyzing and describing linear interassociations of three variables. Extensions of these procedures are available for more than three variables and for regressions of curvilinear form. Furthermore, the applications of the methods presented in this chapter are many and varied and are not by any means limited to the situation of the one illustration given. In general, the most advanced work in application of correlation analysis has been done by those engaged in psychological and education research, and the reader is referred to both the texts and the scientific journals in these fields for elaborations of the methods of this chapter.

SUGGESTED READINGS

Comrie, L. J. (ed.) *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots, and Reciprocals of All Integer Numbers up to 10,000*, 3d. ed. (London: E. and F. N. Spon, 1935).

Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics* (New York: Prentice-Hall, 1939), Chap. 24.

Ezekiel, Mordecai, *Methods of Correlation Analysis*, 2d ed. (New York: Wiley, 1941), Chaps. 9–13.

Johnson, Palmer O., *Statistical Methods in Research* (New York: Prentice-Hall, 1949), Chap. 14.

Kelley, Truman Lee, *The Kelley Statistical Tables*, rev. ed. (New York: Macmillan, 1948).

McNemar, Quinn, *Psychological Statistics* (New York: Wiley, 1949), Chap. 9.

Snedecor, George W., *Statistical Methods: Applied to Experiments in Agriculture and Biology*, 4th ed. (Ames: Iowa State College Press, 1946), Chap. 13.

Stouffer, Samuel A., "Problems in the Application of Correlation to Sociology," *Journal of the American Statistical Association*, 29 (March 1934 Supplement: Proceedings of the American Statistical Association), pp. 52–58.

Yule, G. Udny, and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th ed. (New York: Hafner, 1950), Chap. 12.

Some Uses of Factor Analysis in Sociological Research

THE methods of factor analysis were developed in the field of psychological research, and they have been applied most widely in that field. L. L. Thurstone, the central contributor to this body of method, has seen his contributions become matters of controversy with respect to both their methodological and their interpretative aspects.[1] The existence of such controversies appears to have had the unfortunate effect of limiting exploratory work on the applicability of the methods to other fields.

Coverage of this chapter. It is impossible to treat in one chapter the subject of factor analysis in a comprehensive manner. The authors have nothing to add to either side of still existing controversies. We are indebted to both Thurstone and Harold Hotelling for elucidation and guidance in the field. All we wish to attempt in this chapter is to sketch in broad outline what factor or component analysis attempts to do statistically and to illustrate several types of application to social and economic data.

Brief history. Charles Spearman in 1904 published his first paper on the single common-factor problem and this aspect of factor analysis received attention for several decades.[2] Multiple-factor analysis was introduced by Thurstone in 1931.[3] He considered the single common-factor problem as a special case of the more general formulation. The methods were not very generally available until 1935, when Thurstone published *Vectors of the Mind*. Since that time several texts appeared that presented the methods, and multiple-factor analysis was used by an increasing number of psychologists. Also, mathematical statisticians, such as Hotelling

---

[1] L. L. Thurstone, *Multiple-Factor Analysis: A Development and Expansion of the Vectors of the Mind* (Chicago: University of Chicago Press, 1947).

[2] Charles Spearman, "Correspondence between General Discrimination and General Intelligence," Part III of "General Intelligence Objectively Determined and Measured," *American Journal of Psychology*, XV (1904), pp. 268–272.

[3] L. L. Thurstone, "Multiple-Factor Analysis," *Psychological Review*, XXXVIII (September 1931), pp. 406–427.

and Wilks, contributed to the statistical theory underlying factor analysis.[4]

After 1935 factor analysis was applied to a widening range of phenomena within the areas of research undertaken by psychologists. However, its application in other areas of research has remained quite limited.

WHAT IS FACTOR ANALYSIS?

Factor analysis is a body of methods by which the relationships among a group of variables may be accounted for by a smaller number of variables, or common factors. This is not a complete definition, but probably everyone would agree to this much. Thurstone says, "The factorial methods were developed primarily for the purpose of identifying the principal dimensions or categories of mentality; but the methods are general, so that they have been found useful for other psychological problems and in other sciences as well. Factor analysis can be regarded as a general scientific method."[5]

The most common type of problem to which factor analysis has been applied involves the analysis of data obtained by giving various tests of ability to a group of individuals. If one has the scores of N individuals on n tests, the first step is the computation of all possible pairs of intercorrelations. Since each test will have a correlation with every other test, there will be $\dfrac{n(n-1)}{2}$ correlation coefficients. The matrix of these correlations provides the data on which a factor analysis is performed.

By various operations on the matrix one may identify a set of r factors which account for practically all of the intercorrelations, with r a number often substantially smaller than n, the original number of variables. Regardless of which computation methods are used, the first set of r factors that are identified have the property of being mutually orthogonal or statistically independent—each factor has a correlation of zero with every other factor.

Thurstone then proceeds to rotate these factors—to transform them into another set of r factors that do not necessarily have to be orthogonal. (When the factors are correlated they are nonorthogonal and are referred to as oblique axes.) The object of rotation of the factors, or axes, is to meet the criteria for "simple structure." This means a situation in which

---

[4] (See references at the end of this chapter.) Hotelling and certain others use the term "component analysis" rather than factor analysis. In previous publications the senior author has at times followed Thurstone's usage and at times Hotelling's. In this chapter we are following Thurstone's nomenclature and Hotelling's methods of computation.

[5] Thurstone, *op. cit.*, p. 55.

the individual tests involve some, but generally not all, of the r factors. It is in criteria for rotation of the axes and in the psychological interpretation of results that the controversies have been sharpest. Since the problem of rotation of axes is not encountered in the illustrations to be presented in this chapter and since the problems so far have been enmeshed in controversial aspects of psychological theory, we shall make only brief reference to them.

The set of r orthogonal factors obtained without rotation has interesting properties. They can be considered as summarizing measures of the original group of n variables, excluding the variation in each which is unique to each variable—that is, the part not associated with any other of the variables in the set. In the principal factor solution, which is used in the illustrations given in this chapter, the summarizing nature of the factors is indicated by the fact that the first factor has a greater multiple correlation with the set of original variables than does any other variable. Also, any of the successive factors—say the qth—has a greater multiple correlation with the original variables than does any other variable which is uncorrelated with the first $(q - 1)$ factors. Each factor can be expressed as a linear function of the original n variables, similar to a multiple regression equation, and, in turn, each of the original variables can be expressed as a linear function of the r factors.

On pages 538 to 541 a procedure is given for computing the coefficients, or weights, for the original variables to provide equations for indexes of the factors. The weight of each variable in the equation for a given factor (as obtained by this procedure) is the correlation coefficient between the factor and the variable. Furthermore, the sum of the squares of the coefficients for a given factor divided by the number of variables is the proportion of the variation in the n variables explained by that factor.

In the illustrations that follow factor analysis has been used in one case to obtain an indirect measurement of an item (for which direct measurement is not feasible) from several variables which partially measure or reflect variations in the item. In the other illustrations factor analysis has been used to summarize a large number of variables—to obtain composite measures based on a large number of variables which are more efficient for the purpose at hand than ones obtained by any other method of weighting of the individual variables. These illustrations are limited in scope and are presented as illustrative of the possible applications of factor analysis in the fields of social research in which the varying unit is not one individual person. It is hoped that the use of factor analysis in sociological research will be further explored and extended to a much wider range of types of applications.

CONSTRUCTION OF A LEVEL OF LIVING INDEX

Factor analysis has been used for constructing rural level-of-living indexes for counties of the United States that have been issued by the United States Bureau of Agricultural Economics for different dates.[6] From the data available on a county basis indexes were desired that would reflect geographic differences in the average level of living of farm operator families.[7]

The Census of Agriculture publishes data by counties every five years. From the items available four were selected which would reflect the level of consumption of farm operator families. Previous work had shown that the items chosen were correlated with other items that go to make up the level of living. The items were as follows:

X_1 = percentage of farms with electricity in farm dwelling, 1945

X_2 = percentage of farms with telephone in farm dwelling, 1945

X_3 = percentage of farms with automobiles, 1945

X_4 = mean value of products sold or traded per farm reporting, 1944 (in hundreds of dollars)

Factor analysis was used to determine the weights for combining information on each of these items into a county index. The choice of the method was based on a property of the first factor that is identified in the process of factor analysis. This factor represents the dimension along which the items can best discriminate. It was assumed that the one thing these items could measure best in combination was level of living, since each was a partial measure of level of living and since each was known to be positively correlated with other items in the level of living. Factor analysis provides the weights for combining the items into an index of the first factor.

Table 87 shows the intercorrelations of the four items in the index. The correlations were computed for only a sample of 196 counties drawn from the more than 3,000 counties of the United States, even though indexes were to be computed for all counties.[8]

---

[6] The indexes to be described were published in "Farm Operator Family Level of Living Indexes for Counties of the United States, 1940 and 1945," (Washington: Bureau of Agricultural Economics, May 1947). Some of the material in this section is adapted from two articles that appeared in *Rural Sociology* and is reproduced with the permission of that journal. The articles are "Development of a 1940 Rural Farm Level of Living Index for Counties," 8 (June 1943), pp. 171–180 and "Construction of County Indexes for Measuring Change in Level of Living of Farm Operator Families," *Rural Sociology*, 12 (June 1947), pp. 139–150.

[7] For a discussion of the concept of level of living underlying these indexes see Margaret Jarman Hagood and Louis J. Ducoff, "What Level of Living Indexes Measure," *American Sociological Review*, 9 (February 1944), pp. 78–84.

[8] One reason for using this sample of counties, in addition to reducing computations, was that the results from these counties were processed speedily from the 1945 Census of Agricul-

The correlation matrix shown in Table 87 was then subjected to factor (or component) analysis by the methods to be described in detail in the next section. Only the first factor was identified since no knowledge about the other factors was required to obtain weights for the level-of-living index.

The way in which the first factor is identified is by a set of factor load-

Table 87. INTERCORRELATIONS OF FOUR ITEMS RE-
LATED TO FARM OPERATOR LEVEL OF LIVING, SAMPLE
OF 196 COUNTIES OF THE UNITED STATES, 1945

Item number [a]	Item number [a]			
	1	2	3	4
1				
2	.622			
3	.715	.794		
4	.450	.489	.537	

[a] Identification of items:
 1 = percentage of farms with electricity in farm dwelling, 1945
 2 = percentage of farms with telephone in farm dwelling, 1945
 3 = percentage of farms with automobiles, 1945
 4 = mean value of products sold or traded per farm reporting, 1944 (in hundreds of dollars)

ings—one for each item. These are the correlation coefficients between the first factor and each of the items. In this problem they are as follows:

Correlation with first factor

X_1. 836
X_2. 877
X_3. 920
X_4. 713

Each of the items is highly correlated with the first factor, which we assume to be the level of living. It has been demonstrated that if we wish to use these items to form an index of the first factor, we should weight the items (in standard score form) in proportion to their correlations with the first factor.[9] This seems reasonable as the items more highly correlated with the level of living should receive the higher weights. We, therefore, set up an equation for the index of the level of living, I.

ture to present preliminary results at an early date. The design of this sample also involved factor analysis to be described in the next section.
[9] S. S. Wilks, "Weighting Systems for Linear Functions of Correlated Variables When There is No Independent Variable," *Psychometrika*, 3 (March 1938), pp. 24–43; Harold Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, pp. 417–441, 498–520.

$$I = .836 \left(\frac{X_1 - \bar{X}_1}{s_1} \right) + .877 \left(\frac{X_2 - \bar{X}_2}{s_2} \right) + .920 \left(\frac{X_3 - \bar{X}_3}{s_3} \right)$$

$$+ .713 \left(\frac{X_4 - \bar{X}_4}{s_4} \right)$$

It is not convenient to compute the standard score on each item for 3,000 counties, and so the formula was reduced to the following:

$$I = \frac{.836}{s_1} X_1 + \frac{.877}{s_2} X_2 + \frac{.920}{s_3} X_3 + \frac{.713}{s_4} X_4 -$$

$$\left(\frac{.836}{s_1} \bar{X}_1 + \frac{.877}{s_2} \bar{X}_2 + \frac{.920}{s_3} \bar{X}_3 + \frac{.713}{s_4} \bar{X}_4 \right)$$

From the data on the sample counties, the following means and standard deviations were obtained:

	$\bar{X}$	s
Item 1	49.0	26.0
Item 2	33.3	24.3
Item 3	64.3	24.9
Item 4	31.6	26.0

Substituting these in the formula, we have:

$$I = .0322X_1 + .0361X_2 + .0369X_3 + .0274X_4 - 6.019 \qquad (1)$$

Determining the mean and the zero point of the scale. If the formula just given had been actually used, it would have provided index values for each county with a mean of zero and a standard deviation of approximately 2.5. About half of the counties would have had positive index values, and about half would have had negative values. Such a scale is not in conventional units, and so the scale was transformed. There were two objectives involved in the transformation: (1) setting the United States mean county at 100 and (2) setting a zero point. The transformation was a linear transformation and in no way affected the discriminating capacity of the index. The equation for the transformed index, I', was determined by evaluating the constants a and b in the equation

$$I' = a + bI \qquad (2)$$

Many social and economic indexes have 100 as the value for the average unit or for the base period if it is an index measuring changes over time. Moving the scale from a mean of zero to a mean of 100 could be done without change in size of unit. By adding 100 to the index formula I it could be achieved, but this would not have affected the standard deviation of the distribution of counties with respect to the index. About 67 percent

of the counties would have had values within the range 97.5 and 102.5, which would be highly unconventional. In addition to "sliding" the mean from 0 to 100, it was desirable to "stretch" the index scale so that the variation of counties could be indicated in whole numbers rather than in fractions.

Various ways of stretching the scale could have been chosen. The lowest county could have been set as zero or the standard deviation could have been fixed at some arbitrary number of index points which would not produce negative values.[10] The choice of a method was affected by a second objective of constructing the index. An index was needed not only to measure differences among counties at a given point of time, but also to measure changes over time for the same county. To serve the latter objective percentage rates of change as well as index points of change were required.

A percentage difference in scale or index points has no meaning unless the scale has some meaningful zero point. For example, suppose that on one index scale County A changed from a value of 40 to a value of 60 during a five-year period, an increase of 50 percent. But if the scale has no real zero point, we can slide its mean upward by adding 100 without affecting the differentiating capacity of the scale. On the transformed scale County A would show a gain from 140 to 160 of only 14 percent. Neither percentage means anything unless there is a real zero point and a defined scale unit.

In the present example the value of zero was taken to represent a bare subsistence level of living at which the values for a county on each of the items would be zero. Farm operators in a county with an index value of zero would have a total absence of cash agricultural income, modern household facilities, and automobiles. With the value of the mean already to be 100, setting of the zero point defines the unit of the level-of-living index scale. One unit of the scale is one one-hundredth of the difference between the average level of living of farm operator families in the average county of the United States in 1945 and a subsistence level in a hypothetical county of farm operators without any cash farm income, household conveniences, or modern transportation.

We next determine the values for a and b that will slide our scale and stretch it as desired. The mean of the present scale I is zero, and we want the mean of I′ to be 100. Substituting these values in (2), we have

$$100 = a + b\,(0)$$
$$a = 100$$

---

[10] The former method was used in the illustration on regional delineation given later in this chapter; the latter was used by the Bureau of Agricultural Economics in rural level-of-living indexes published in 1943.

To get the value of b, we substitute values when all of the items are zero, the zero point desired for I'.

$$0 = 100 + b\,(-6.019)$$
$$b = 16.71$$

The transformation equation is

$$I' = 100 + 16.71\ I \tag{3}$$

Substituting the right side of equation (1) for I, we have the transformed equation,

$$I' = .538X_1 + .603X_2 + .617X_3 + .460X_4 \tag{4}$$

Index formula (4) was actually applied to the 1945 data for each county of the United States. To get comparable indexes for other years an adjustment was made to the coefficient of X_4 to allow for the changing amount of goods that the farmer's dollar would buy.[11] Map 2, page 55 shows the counties of the United States grouped by quintiles according to their rank on the 1945 index.

FACTOR ANALYSIS INDEXES FOR STRATIFICATION IN SAMPLING[12]

The purpose of stratification in sampling is to utilize available information on the units to be sampled in such a way as to assure better representation with respect to certain characteristics of the units than would be expected from simple random sampling. The process of stratification involves, first, the choice of "control" characteristics and, second, some method of utilizing information on the control characteristics which

---

[11] For the formulas for other years see *Farm Operator Family Level of Living Indexes for Counties of the United States, 1930, 1940, 1945, 1950* (Washington: Bureau of Agricultural Economics, December 1951).

[12] This section is reproduced with minor modifications from Margaret Jarman Hagood and Eleanor H. Bernert, "Component Indexes as a Basis for Stratification in Sampling," *Journal of the American Statistical Association,* 40 (September 1945), pp. 330–341, with the permission of the editors of that journal.

In addition to the 70-county sample described, other national samples were developed from similar use of the same index values for counties. These included a general purpose sample of 101 counties used by the Bureau of Agricultural Economics for its Quarterly Survey of Agriculture in 1945, a special-purpose sample of 158 counties used by that bureau in 1945 and 1946 for recurring surveys of farm wage rates, and a 223-county sample used by the Census of Agriculture for early processing of results from both the 1945 and the 1950 Censuses of Agriculture. For a fuller description of the sample designs used in these surveys see "Technical Supplement to the Bureau of Agricultural Economics General Purpose 101—County Sample," (Washington: Bureau of Agricultural Economics, 1945, *Mimeographed*) and Louis J. Ducoff and Margaret Jarman Hagood, "Wages and Wage Rates of Hired Farm Workers in the United States. United States and Major Geographic Division, March 1945," *Surveys of Wages and Wage Rates in Agriculture,* Report No. 4, (Washington: Bureau of Agricultural Economics, 1945).

will group the units to be sampled into strata, each containing units relatively homogeneous with respect to the control characteristics. Both nonquantitative classifications, such as geographic location, and quantitative variables may be used as control characteristics for stratification.

The criterion for choice of control characteristics is their relation to the item on which observations are to be made in the sample for the purpose of making estimates for the universe. The more closely a single control characteristic is correlated with the item to be estimated from the sample, the greater will be the improvement in efficiency of estimation from a sample stratified to assure representativeness with respect to that control characteristic over a sample drawn without any stratification.

Several considerations complicate the problem of choice of control characteristics. After one control has been chosen, the simple criterion of highest correlation with the item to be estimated has to be modified in choice of the second and subsequent controls, since the interrelationships of the control characteristics have to be taken into account. In the straightforward case where a single item is to be estimated from the sample, and its correlations with available control characteristics are known, along with their intercorrelations, the choice of successive control characteristics is similar to the choice of successive estimating variables in a multiple regression problem.

In the more complex and far more common case in social and economic surveys estimates are to be derived from the sample not for one but for many items. Moreover, often no precise information is available as to the degree of correlation of the various items to be estimated from the sample with the characteristics on which data are available for possible use as controls. In such a case judgments of persons familiar with the subject matter may have to be relied on for determining which of a group of possible controls are likely to be most closely correlated with the items to be estimated from the sample.

When control characteristics have been selected, the next problem is to develop a method of utilizing information on the selected controls which will group the units into strata containing units as similar as possible with respect to the control characteristics. Indexes derived from factor analysis of the intercorrelations of control characteristics which are quantitative variables have been used as a technique for stratification.

For certain field work of the Bureau of Agricultural Economics a relatively small sample of counties was desired for use in continuing and periodic social and economic studies relating to the attitudes, folkways, institutions, population movements, and other behavior patterns of farm people. Because certain types of studies involve communities and other units larger than a single farm or farm family and because administrative

considerations limited the number of different locations that could be studied, the county was chosen as the primary sampling unit, and the number of counties to be included in the sample was set at approximately 70.

Major type-of-farming regions of the United States were chosen as the primary basis of stratification. Geographic differences condition the type of agriculture prevailing in the different regions of the United States and, thus, affect importantly the social and economic characteristics of farm people. In addition, it was desired that the sampling design provide a basis for summaries for major type-of-farming regions as well as for the United States. The major type-of-farming regions were adapted from a revision of the well-known BAE type-of-farming delineation developed in 1937. All counties of the United States were thus first grouped into seven major type-of-farming regions and a residual group for which no regional summary would be attempted but which would be necessary to supplement the seven major type-of-farming region samples in making national summaries.

Although geographic and type-of-farming criteria were considered the most important single basis for stratification and were used as the primary basis, geographic stratification per se was not used further in determining strata within major type-of-farming regions. This is a departure from the more customary practices of sampling in rural social surveys. Generally, within a major geographic region or state finer strata have been developed by the use of minor type-of-farming sub-regions consisting of contiguous counties. Such a sample design, however, relies wholly on geographic control as indicated by the subregional classification as the only basis for stratification.

The hypothesis underlying the development of the plan presented is that the social and economic characteristics of farm people are importantly influenced by factors other than geographic; hence, information relating to these other factors should also be used in stratification in order to assure satisfactory representation of the variations in these factors. However, purely geographic or type-of-farming area control may be preferable for studies of phenomena, such as crop yields, which within a given region or state are more completely determined by physiographic factors (such as soil type, rainfall, topography, etc.).

From the range of characteristics available for use as controls in delineating strata within each major type-of-farming region, some 12 to 14 variables were chosen from the 1940 Censuses of Agriculture and Population and related sources. The considerations governing the choice of the variables will not be detailed here, but, in general, the most important criterion was relevance to, or estimated correlation with, the broad classes of phenomena to be observed in the sample counties.

A set of control variables was selected for each major type-of-farming region separately with certain variables used in every region. Generally, there was more uniformity in the group of variables relating to the farm population for the several regions than among the variables relating to agricultural characteristics. For example, the proportion of the total population of the county which is rural-farm, a rural-farm level-of-living index, and information on migration for the 1930–1940 decade and on total population change for the period 1940–1943 were used as control variables in every region. Among the agricultural characteristics mean age of farm operators extent of off-farm work of farm operators and some variable relating to hired farm labor were used in almost every region, while the variables relating to type of production, size of farm, mechanization, etc. were selected according to their importance in any given region. As examples, in the Corn Belt the proportion of farm land in corn was an important agricultural control, in the Cotton Belt the percentage of farms operated by sharecroppers, in the Dairy Region the number of cows milked per farm reporting, in the General and Self-Sufficing Region the proportion home consumption comprised of the total value of agricultural production, in the Range Livestock Region the number of cattle kept principally for beef, in the Wheat Belt the value of implements and machinery per farm, and in the Western Specialty Crop Areas the proportion of the county's total agricultural production produced on farms with a value of products of $10,000 or more.

For illustration of the use of factor indexes in stratification of counties for sampling, data and procedures are presented for the General and Self-Sufficing Region. This region extends from New England down through the Appalachians into the northern edge of Georgia and westward to the Ozarks. The region includes 19 percent of the nation's farm population. Because the major criterion in delineation of the General and Self-Sufficing Region was the absence of any dominant crops or livestock in agricultural production, the agricultural characteristics chosen for controls were more general than in the case of the other regions. The 12 population and agricultural variables used as controls in the region are listed in Table 88.

From the 552 counties in the General and Self-Sufficing Region only about 10 were to be included in the national sample. The problem of stratification then was one of grouping the 552 counties into 10 groups or strata, with each stratum containing counties as nearly alike as possible with respect to the 12 characteristics which had been chosen as control variables.

A further requirement in making the stratification was imposed, namely, that each of the 10 strata should contain approximately the same number of farm people. Thus, a regional summary of 10 cultural studies made in the 10 sample counties would not be distorted by inequality in the

importance of the several studies, as would be the case if one county represented two or three times as many people as another.

To utilize information on all the selected control variables in stratification, mutually uncorrelated factor indexes were employed. Each index is a linear function of the 12 control variables, with the weights for

Table 88. INTERCORRELATIONS OF 12 POPULATION AND AGRICULTURAL VARIABLES FOR 552 GENERAL AND SELF-SUFFICING COUNTIES

Identification number of variable	Identification number of variable											
	1	2	3	4	5	6	7	8	9	10	11	12
1	—											
2	.179	—										
3	.305	.300	—									
4	−.399	−.115	−.153	—								
5	−.600	−.228	−.765	.234	—							
6	−.260	.089	.033	.150	.067	—						
7	−.510	−.195	−.699	.219	.672	.089	—					
8	−.440	.152	.216	.260	−.080	.416	.270	—				
9	.357	.378	.773	−.195	−.815	−.121	−.494	.386	—			
10	−.468	−.207	−.441	.310	.695	.037	.353	−.135	−.644	—		
11	.409	.227	.318	−.452	−.421	−.026	−.445	−.154	.455	−.559	—	
12	.256	.055	−.254	−.230	.083	−.184	.010	−.350	−.284	−.048	.139	—

Identification of variable

1 Percent of total population which is rural-farm, 1940
2 Altitude measure for County Seat
3 Replacement rate of rural-farm males, 25–69 years of age, 1940
4 Percent change in total civilian population, 1940–43
5 Rural-farm level-of-living index, 1940
6 Percent change in rural-farm population through migration, 1930–1940
7 Mean age of farm operator, 1940
8 Percent farm operators reporting 100 days or more of work off farms comprise of those reporting work off farms, 1939
9 Value of home consumption as percent of value of all farm products, 1939
10 Mean value of all farm products, 1939
11 Family labor as percent of all farm labor, 1940
12 Value of livestock sold as percent of value of all farm products, 1939

the variables being determined by factor analysis of the matrix of intercorrelations of the variables. Factor analysis of the matrix shown in Table 88 yielded the correlation coefficients between the factors and each of the variables shown in Table 89. These coefficients of Table 89 formed the weights used in constructing indexes of the factors.

If the problem were to select a single variable for grouping counties into strata, the first factor is the variable which would afford the "best"

Table 89. CORRELATION OF 12 POPULATION AND AGRICULTURAL VARIABLES WITH THREE INDEPENDENT FACTORS, 552 GENERAL AND SELF-SUFFICING COUNTIES

Identification number of variable	Factor			Proportion of variation explained by three factors
	I	II	III	
				(Percent)
1	−.669	−.499	.058	70.0
2	−.388	.168	.458	38.9
3	−.792	.357	−.247	81.6
4	.430	.447	−.322	48.8
5	.906	−.127	.158	86.2
6	.127	.515	.471	50.3
7	.753	.119	.327	68.8
8	.006	.866	.262	81.9
9	−.847	.388	−.064	87.2
10	.759	−.042	−.244	63.7
11	−.660	−.228	.359	61.6
12	.023	−.659	.391	58.8
Proportion of variation in 12 variables explained by factor (percent)	37.9	19.1	9.6	66.6

Identification of variable
 1 Percent of total population which is rural-farm, 1940
 2 Altitude measure for county seat
 3 Replacement rate of rural-farm males, 25–69 years of age, 1940
 4 Percent change in total civilian population, 1940–1943
 5 Rural-farm level-of-living index, 1940
 6 Percent change in rural-farm population through migration, 1930–1940
 7 Mean age of farm operator, 1940
 8 Percent farm operators reporting 100 days or more of work off farms comprise of those reporting work off farms, 1939
 9 Value of home consumption as percent of value of all farm products, 1939
10 Mean value of all farm products, 1939
11 Family labor as percent of all farm labor, 1940
12 Value of livestock sold as percent of value of all farm products, 1939

basis for grouping together counties so as to provide maximum homogeneity within groups with respect to the 12 control variables.[13] However, we do not have direct information on the counties' values with respect to the first factor, which could be used in grouping the counties into strata. But the factor index as described above affords a basis for estimating for each county its value with respect to the first factor.

The index equation for estimating a county's value for the first factor (I) from its value on the 12 variables listed in Table 88 is as follows:

---

[13] See references in footnote 9 of this chapter.

$$I = -.669z_1 - .388z_2 - .792z_3 + .430z_4 + .906z_5 + .127z_6$$
$$+ .753z_7 + .006z_8 - .847z_9 + .759z_{10} - .660z_{11} + .023z_{12},$$

where the z's represent standard scores of the control variables. This index was computed for each of the 552 counties in the region. On the basis of the resulting index values, the counties were grouped into five classes, each containing approximately the same number of farm population. Examination of the index formula indicates that counties with a high value on Index I tend to have a high value on the rural-farm level of living index ($r = .906$), a high total value of products per farm ($r = .759$), a high average age of farm operators ($r = .753$), and low values on the proportion home consumption is of all farm products ($r = -.847$), on the replacement rates of males 25–69 years of age ($r = -.792$), on the proportion of population living on farms ($r = -.669$), and on the proportion family workers comprise of all farm workers ($r = -.660$). Thus, classification of counties into five groups according to this index means that the counties in a group will be relatively similar with regard to the variables cited which are highly correlated (either positively or negatively) with the first factor.

To arrive at 10 strata it would have been possible to use 10 class intervals of Index I. Instead, a second factor index, uncorrelated with Index I, was developed, and each of the five groups determined by Index I was subdivided on the basis of counties' values on Index II. Because the factors are mutually independent, the control attained by use of the second factor is a net addition to the control attained by use of the first factor. The correlation coefficients in the original matrix were reduced by the amount explained in each by the first factor. For this reduced matrix, a repetition of the processes used in deriving the weights for the first factor yielded weights to construct an index for the second factor. Index values were computed for each county, and the five groups of counties were subdivided into 10 strata according to county values on Index II. From Table 89 it can be seen that the proportion of farm operators working 100 days or more off the farm during a year is most highly correlated with Index II, and the importance of livestock next most highly (negatively).

In a similar manner six to 12 strata were delineated for each of the seven major type-of-farming regions and for the residual group of counties not included in any major type-of-farming region. Thus, all of the counties of the United States (except 14 which have no rural-farm population) were grouped into 70 strata, with each stratum including counties relatively homogeneous with respect to some 12 agricultural or farm population characteristics deemed important for that region.

The total number of counties to be studied, approximately 70, was set largely by administrative considerations. Since it was desired to maxi-

mize the control afforded by stratification, the number of strata to be delineated was also set at approximately 70, with only one county from each stratum to be included in the sample. Allocation of the 70 strata among the seven major regions and the residual group was made by a first approximation of equal numbers to each region, modified in the light of: (1) variability of counties within the region, (2) the region's share of farm population, and (3) the region's share of agricultural production.

When the number of strata for a region was set, the next problem was to determine how many factor indexes to use and how many class intervals of each. The principles generally followed in the present stratification were: (1) to use an index for each factor explaining more than 10 percent of the total variation of the control variables; (2) to use more class intervals for the index of the factor which explained the greatest proportion of the variation of the control variables. For example, in the General and Self-Sufficing Region only the first two factors (explaining 37.9 and 19.1 percent, respectively) were used, since the third factor explained less than 10 percent of the total variation. More class intervals of the index of the first factor were used than of the second because it explained approximately twice as great a percentage of the variation; however, there was no precise guidance on the best allocation of classes between the two.

During the development of this sample design, some experimental work was done to afford guidance at different steps. However, pressure of time did not permit full exploration of the relative advantages of alternative methods or procedures. With the increasing use of sampling in national surveys, governmental and private, there is great need for further research in determining which methods of stratification are most efficient for given survey situations. In further exploration of the problems of stratification in sampling, the following questions need to be worked on:

(1) For what types of inquiries is the combination of geographic control and other control variables more or less efficient than only geographic control for stratification?

(2) If control variables are to be used, when is it more efficient to stratify by class intervals of one, two, or three observed variables and when is it more efficient to stratify by one, two, or three factor indexes based on information from a considerably larger number of observed variables?

(3) If factor indexes are to be used, what criteria can be used in determination of the number of control variables to be used, the number of factor indexes to be used, and for each index, the number of class intervals?

Computation procedures for factor analysis. The computation procedures used in this problem are summarized in the following instructions and computation tables.

Table 90. PROCEDURES FOR COMPUTING INDEX WEIGHTS BY FACTOR ANALYSIS

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1	1.000	.179	.305	-.399	-.600	-.260	-.510	-.440	.357	-.468	.409	.256
2	.179	1.000	.300	-.115	-.228	.089	-.195	.152	.378	-.207	.227	.055
3	.305	.300	1.000	-.153	-.765	.033	-.699	.216	.773	-.441	.318	-.254
4	-.399	-.115	-.153	1.000	.234	.150	.219	.260	-.195	.310	-.452	-.230
5	-.600	-.228	-.765	.234	1.000	.067	.672	-.080	-.815	.695	-.421	.083
6	-.260	.089	.033	.150	.067	1.000	.089	.416	.386	.037	-.026	-.184
7	-.510	-.195	-.699	.219	.672	.089	1.000	.270	-.494	.353	-.445	.010
8	-.440	.152	.216	.260	-.080	.416	.270	1.000	.386	-.135	-.154	-.350
9	.357	.378	.773	-.195	-.815	.386	-.494	.386	1.000	-.644	.455	-.284
10	-.468	-.207	-.441	.310	.695	.037	.353	-.135	-.644	1.000	-.559	-.048
11	.409	.227	.318	-.452	-.421	-.026	-.445	-.154	.455	-.559	1.000	.139
12	.256	.055	-.254	-.230	.083	-.184	.010	-.350	-.284	-.048	.139	1.000
Σ1	-.171	1.635	.633	.629	-.158	1.290	.270	1.541	.796	-.107	.491	.193
Σ2	-.250	1.484	1.255	.352	-.038	1.285	-.299	1.815	1.463	-.738	.509	-.616
Σ3	.120	1.664	2.213	.042	-1.837	1.211	-1.046	2.078	2.447	-1.597	.998	-1.204
Σ4	1.309	1.964	3.508	-.662	-1.796	.775	-2.355	1.865	3.807	-.281	2.264	-1.326
Σ5	1.663	1.704	3.197	-1.006	-3.050	.207	-2.532	.952	3.336	-2.367	2.066	-.878
Σ6	2.672	1.979	3.953	-1.638	-4.253	-.161	-3.403	.644	4.207	-3.482	2.873	-.640
Σ7	-3.074	-1.993	-4.053	1.925	4.508	.402	3.659	-.335	-4.333	3.745	-3.150	.396
Σ8	-3.206	-1.962	-4.004	2.034	4.517	.517	3.708	-.153	-4.283	3.771	-3.220	.235
Σ9	-3.277	-1.952	-3.986	2.094	4.527	.578	3.739	-.061	-4.256	3.786	-3.261	.184
Σ10	-3.314	-1.947	-3.977	2.125	4.533	.609	3.756	-.013	-4.253	3.795	-3.282	.148
Σ11	-3.327	-1.945	-3.972	2.139	4.534	.621	3.762	.005	-4.249	3.799	-3.294	.133
Σ12	-3.342	-1.946	-3.974	2.149	4.540	.631	3.771	.020	-4.250	3.804	-3.301	.122
Σ13	-3.347	-1.946	-3.972	2.152	4.540	.635	3.773	.027	-4.248	3.805	-3.303	.117
Σ14	-3.350	-1.946	-3.972	2.155	4.542	.638	3.774	.031	-4.249	3.806	-3.305	.115
Σ15	-3.351	-1.945	-3.971	2.156	4.541	.638	3.775	.033	-4.247	3.805	-3.306	.113

Table 91

Variable	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}*	w_{11}	w_{12}	w_{13}	w_{14}	w_{15} (w_a)	$a_a = (.906 \times w_a)$
1	−.105	−.138	.049	.344	.499	−.628	−.682	−.710	−.724	−.727	−.734	−.736	−.737	−.738	−.738	−.669
2	1.000	.818	.680	.516	.511	−.465	−.442	−.434	−.431	−.429	−.429	−.429	−.429	−.429	−.428	−.388
3	.387	.691	.904	.921	.958	−.929	−.899	−.886	−.880	−.876	−.876	−.875	−.875	−.875	−.874	−.792
4	.385	.194	.017	−.174	−.302	.385	.427	.450	.463	.470	.472	.473	.474	.474	.475	.430
5	−.097	−.021	−.751	−.472	−.914	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.906
6	.789	.708	.495	.204	.062	.038	.089	.114	.128	.134	.137	.139	.140	.140	.140	.127
7	.165	−.165	−.427	−.619	.759	.800	.812	.821	.826	.829	.830	.831	.831	.831	.831	.753
8	.943	1.000	.849	.490	.285	−.151	−.073	−.034	−.013	−.008	.001	.004	.006	.007	.007	.006
9	.487	.806	1.000	1.000	1.000	−.989	−.961	−.948	−.940	−.935	−.937	−.936	−.936	−.935	−.935	−.847
10	−.065	−.407	−.653	−.074	−.710	.819	.831	.834	.836	.838	.838	.838	.838	.838	.838	.759
11	.300	.280	.408	.595	.619	−.676	−.699	−.713	−.720	−.726	−.727	−.727	−.728	−.728	−.728	−.660
12	.118	−.339	−.492	−.348	−.263	.150	.088	.056	.041	.034	.029	.027	.026	.025	.025	.023

* On the basis of the trend indicated by the preceding weights, these values were estimated in order to obtain convergence more quickly.

Table 92

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1	.552	−.081	−.225	−.111	.006	−.175	−.006	−.436	−.210	.040	−.033	.271
2	−.081	.849	−.007	.052	.124	.138	.097	.154	.049	.087	−.029	.064
3	−.225	−.007	.373	.188	−.047	.134	−.103	.221	.102	.160	−.205	−.236
4	−.111	.052	.188	.815	−.156	.095	−.105	.257	.169	−.016	−.168	−.220
5	.006	.124	−.047	−.156	.179	−.048	−.010	−.085	−.048	.007	.177	.062
6	−.175	.138	.134	.095	−.048	.984	−.007	−.010	−.013	−.059	.058	−.187
7	−.006	.097	−.103	−.105	−.010	−.007	.433	.265	.144	−.219	.052	−.007
8	−.436	.154	.221	.257	−.085	−.010	.265	1.000	.391	−.140	−.150	−.350
9	−.210	.049	.102	.169	−.048	−.013	.144	.391	.283	−.001	−.104	−.265
10	.040	.087	.160	−.016	.007	−.059	−.219	−.140	−.001	.424	−.058	−.065
11	−.033	−.029	−.205	−.168	.177	.058	.052	−.150	−.104	−.058	.564	.154
12	.271	.064	−.236	−.220	.062	−.187	−.007	−.350	−.265	−.065	.154	.999

1. To the matrix of intercorrelations shown in Table 88, add a "1" in each diagonal cell to obtain Table 90.

2. Add the correlation coefficients in each column of Table 90 obtaining the row of sums shown in line designated as $\Sigma 1$ at bottom of table.

3. Divide each sum shown in line $\Sigma 1$ by the largest single sum (1.635). Enter the quotients in the first column of Table 91, labeled w_1. Note that the weight derived from the sum of the kth column of Table 90 is entered in the kth row of the column labeled w_1 in Table 91. (The rows of Table 91 should be spaced exactly the same as the rows of Table 90.)

4. Fold back Table 91 at the right edge of column w_1 and place it on Table 90 so that the entries of column w_1 in Table 91 (which must be on a separate sheet) are just to the left of the entries of column 1 in Table 90. Cumulate the sum of the products of the entries in each row of the two columns—that is, $-.105 \times 1.000 + 1.000 \times .179 + \cdots + .118 \times .256 = -.250$. The individual products do not have to be written down, as only their sum is needed and this can be accumulated in the appropriate dials of the calculator. Enter the sum in column 1 of the line designated as $\Sigma 2$ underneath Table 90.

5. Move Table 91 over to the right by one column so that column w_1 of Table 91 will be just to the left of column 2 of Table 90. Cumulate the sum of the products of the entries of the two columns for each row and enter this sum in the second column of row $\Sigma 2$.

Continue as above for each successive column of Table 90, obtaining an entry for each column of the row $\Sigma 2$.

6. Divide each entry in row $\Sigma 2$ of Table 90 by the largest single entry in that row. Unfold Table 91, and enter the quotients in column w_2 of Table 91 (with the entry obtained from the kth column of row $\Sigma 2$ placed in the kth row of column w_2).

7. Fold back Table 91 at the right edge of column w_2 and place it on Table 90 so that the entries of column w_2 in Table 91 are just to the left of the entries in column 1 of Table 90. As before cumulate the sum of the products of the pair of entries in each row and enter the sum in column 1 of row $\Sigma 3$.

Move Table 91 over to the right by one column and cumulate products to obtain the entry in column 2 of row $\Sigma 3$.

Continue as above for each successive column.

8. Continue to repeat instructions 6 and 7 for additional Σ rows in Table 90 and additional w columns in Table 91 until a set of weights is obtained in the w column identical with (or within .001) the set obtained in the preceding column. (These weights will not change with further iteration; since they are "stable" weights, the column is designated as w_s.)

9. The weights in column w_s are proportional to the correlation coefficients of the first principal factor with the 12 variables. To obtain the actual correlation coefficients multiply each weight by the following quantity:

$$\sqrt{\frac{\text{Largest single sum in last row of sums in Table 90}}{\text{Sum of the squares of the weights in row } w_s \text{ of Table 91}}}$$

Enter the products in column a_s of Table 91. These are the correlation coefficients of the first factor with each of the variables, as shown in Table 89,

and will be identified as a_1, a_2, etc. for the several variables. (The sum of the squares of the a_i's divided by the number of variables is the proportion of variation in the 12 variables explained by the first factor.)

10. To begin work on the second factor, Table 92 must be prepared. It is the reduced matrix of intercorrelations of the variables with any correlation explained by the first factor removed. To get an entry for a cell ij of Table 92, take the corresponding entry of Table 90 and subtract from it the product of $a_i a_j$. The entry in column 1, row 1 of Table 92 is obtained thus:

$$1.000 - (-.669)^2 = .552$$

The entry in column 1, row 2, thus:

$$.179 - [(-.669)(-.388)] = -.081.$$

11. Proceed to obtain the correlation coefficients of the second factor with the variables by operating on Table 92 exactly as described in steps 2–9 for Table 90. (A new table 93 corresponding to Table 91 will have to be prepared for recording the weights, etc. for the second factor.) Designate the correlation coefficients of the second factor with the variables as b_1, b_2, etc.

12. If further factors are required, follow instruction 10 to reduce again the previously reduced matrix and then repeat steps 2–9.

FACTOR ANALYSIS INDEXES FOR REGIONAL DELINEATION [14]

In regional or subregional delineation the objective of grouping together homogeneous units is the same as in the problem of stratification, but, in addition, there is usually the requirement that the units grouped together be contiguous. The example presented in this section involves the grouping of states into major regions of the United States, but the methods could be applied to counties or other geographic units.

The problem illustrated here was one of dividing the United States into some six to a dozen groups of contiguous states with each group of states as internally homogeneous as possible with respect to a large group of items taken from the 1940 Censuses of Population and Agriculture and vital statistic reports. As a first step, the items were grouped according to the following subjects:

> Agriculture (52 items)
> Land use (4 items)
> Crops (12 items)
> Livestock (10 items)
> Farm tenure (11 items)
> Farm values (8 items)
> Farm finance (7 items)

---

[14] This section is adapted from Margaret Jarman Hagood, "Statistical Methods for Delineation of Regions Applied to Data on Agriculture and Population," *Social Forces*, 21 (March 1943), pp. 288–297, with the permission of the editors of *Social Forces*.

Population (52 items)
Residence (6 items)
Race (5 items)
Sex (6 items)
Age (6 items)
Education (5 items)
Employment status (5 items)
Occupation (8 items)
Vital statistics (7 items)

For each of these 14 groups the intercorrelations among the items in the group were computed for the 48 states. Next, a factor analysis was made of each of the 14 correlation matrices to identify the first factor loadings. From these factor loadings an index formula was constructed for each group by methods explained in the preceding sections. These 14 index formulas were evaluated for each of the 48 states from data on the 52 agricultural and 52 population items.

When the states were mapped according to their values on the various group indexes, the states with similar index values tended to form contiguous groupings, but this was not invariably so. For example, Map 5 shows that with respect to the educational status index, the states in the Far West and those in the extreme Northeast fell in the same interval. (On this index scale, states with high educational status have the lower values.)

In order to use information on all the 104 items simultaneously, a second level of index construction was used. Correlations were computed for all pairs of the 14 group indexes, as is shown in Table 93. A factor analysis of this matrix produced the formula shown at the bottom of the table for a composite index.

Regions could be delineated by assigning to the same regions states which fall into the same class interval on the composite index. However, states in one interval might not always form a contiguous grouping, a condition necessary for regions as we have set the problem.

"Coefficients of similarity." To supplement the use of the composite index in the grouping of states into regions, we need to make what might be called a case analysis of each state with regard to its neighbors. For instance, it makes no difference in the allotting of states to regions whether or not California is like New Jersey, but it does make a difference whether it is more like Arizona or Nevada. For California cannot be put into the same region as New Jersey because of geographical separation, while it might be put in the same region with either, both, or neither of its neighbors mentioned.

Hence, a supplementary technique was devised for affording a measure of degree of similarity between what we shall term the "agricultural-

Map 5. Educational Status Index, 48 States, 1940. (Adapted by the Institute for Research in Social Science, University of North Carolina, from Goode's Base Map Series, Copyright 1938. By permission of the University of Chicago Press.)

INDEX VALUES

0-13
14-22
23-43
44-100

Table 93. MATRIX OF INTERCORRELATIONS OF ITEMS FOR COMPOSITE AGRI-
CULTURE-POPULATION INDEX, 48 STATES, 1940

	Aa	Ab	Ac	Ad	Ae	Af	Pa	Pb	Pc	Pd	Pe	Pf	Pg	Ph
								Components						
Aa														
Ab	.752													
Ac	−.318	−.675												
Ad	−.270	.204	−.614											
Ae	−.099	−.465	.819	−.695										
Af	−.287	−.542	.483	−.613	.688									
Pa	.236	−.079	.558	−.505	.608	.114								
Pb	.335	.555	−.508	.574	−.519	−.805	−.036							
Pc	−.609	−.476	.125	−.067	.189	.629	−.452	−.531						
Pd	−.029	.382	−.743	.784	−.834	−.569	−.715	.443	−.069					
Pe	.262	.557	−.682	.728	−.676	−.688	−.369	.621	−.357	.768				
Pf	−.462	−.295	−.191	.104	−.197	.360	−.642	−.437	.690	.327	−.149			
Pg	−.468	−.637	.579	−.556	.578	.778	.162	−.757	.528	−.496	−.760	.457		
Ph	−.268	.189	−.556	.927	−.689	−.582	−.536	.477	−.112	.856	.742	.155	−.475	

Equation for standard measures:

Composite index = $.308z_{Aa} + .728z_{Ab} - .924z_{Ac} + .923z_{Ad} - .971z_{Ae} - .943z_{Af} - .504z_{Pa} + .866z_{Pb}$
$- .455z_{Pc} + .958z_{Pd} + 1.000z_{Pe} - .116z_{Pf} - .926z_{Pg} + .902z_{Ph}$.

Identification of components:

Aa—Land use index (4 items)
Ab—Crop index (12 items)
Ac—Livestock index (10 items)
Ad—Farm tenure index (11 items)
Ae—Farm values index (8 items)
Af—Farm finance index (7 items)
Pa—Residence composition index (6 items)
Pb—Race composition index (5 items)
Pc—Sex composition index (6 items)
Pd—Age composition index (6 items)
Pe—Educational status index (5 items)
Pf—Employment status index (9 items)
Pg—Occupational composition index (8 items)
Ph—Vital statistics index (7 components)

Source: *Sixteenth Census of the United States, 1940*, Agriculture and Population bulletins; Vital Statistics Special Reports, Vol. 10.

population profiles" of states. First all the 104 series of variables used in the 14 group indexes were put into a standard coded form so that the 48 states would have a mean of 50 and a standard deviation of 10 on each series. Next for any pair of states where the degree of resemblance was of interest the correlation between the two states was computed by considering the sequence of 104 coded measures for one state as one variable, and the sequence of 104 measures for the other state as the other variable.

Delineation of regions. All the data on agriculture and population considered have been condensed into the numerals shown on Map 6. Each circled numeral shows the state's value on the composite agricultural-population index (scaled so that the lowest state will have a value of zero and the highest a value of 100); each small numeral shows

Map 6. Composite Agriculture-Population Index Values and Coefficients of Similarity in Agriculture-Population Profiles, 48 States, 1940. (Adapted by the Institute for Research in Social Science, University of North Carolina, from Goode's Base Map Series, Copyright 1938. By permission of the University of Chicago Press.)

the coefficient of correlation between the two states indicated by lines drawn to state names. We are now ready to delineate a set of major regions from this synthesis of information. The criteria for delineation shall be as follows: regions shall be geographically contiguous, they shall be as few in number as possible to include only relatively homogeneous states, they shall consist of states with as nearly similar composite agricultural-population index values as possible, they shall consist of states with as high intercorrelations as possible, they shall not consist of one state alone.

Delineation is most easily done in two steps: first, the easy determination of regional nuclei of states which are without question so homogeneous that they should be in the same region; and second, the allocation of the remaining states to these nuclei and the possible combination of the nuclei in such a way as to compromise least with the criteria just formulated. The states in the Southeast (as delineated in *Southern Regions of the United States*), with the exception of Florida, form one such nucleus; Maine, Vermont, and New Hampshire another; Idaho, Montana, and Wyoming another; while certain states like Missouri are not very similar to any of their neighbors. Ten such nuclei of two or more states can be identified, leaving 12 states, each of which is to be allotted to the most similar nucleus. In some cases there is no problem of judgment; Florida has to be allotted to the Southeast if there are to be no one-state regions. On the other hand, the allocation of Missouri is a difficult matter involving judgment. At this stage other information such as historical data might well be utilized along with the summary of quantitative material in determining the allocation of "problem" states.

No one delineation is uniquely determined by this method. For a number of regions limited to six, the data of Map 6 indicate the same delineation used in *Southern Regions of the United States* except in the case of West Virginia, which falls with the Southeast rather than with the Northeast. An alternate possible delineation of seven regions divides the Northeast into two parts—New England, with Maine, New Hampshire, and Vermont; and the Industrial Northeast, with the remaining eight states; adds Oklahoma and Texas to the Southeastern states; and groups Nevada with Arizona and New Mexico, leaving only the states actually bordering the Pacific in the Far West.

CURRENT AND FUTURE DEVELOPMENTS IN FACTOR ANALYSIS

Thurstone believes that a profitable line of future work will be in the direction of developing nonmetric methods of factor analysis for dealing with scores that are in the form of ranks.[15] He wrote in 1945 "The main

---

[15] Thurstone, *op. cit.*, p. xiv.

problem for factor analysis is the more fundamental one of charting the alternative factor patterns in new fields. To develop this subject with more powerful nonmetric methods is future work that may leave our present efforts obsolete." [16]

Louis Guttman, Scientific Director of the Israel Institute of Applied Social Research, presented lectures in several universities of the United States in the spring of 1951 on a new approach to factor analysis. He is formulating a more general theory of factor analysis, which embraces Thurstone's multiple factor theory, his own work on scaling, and Paul Lazarsfeld's work on latent attributes as special cases.

Also, work is going on by mathematical statisticians aimed toward developing the sampling distribution of the factor loadings that would permit criteria to be set objectively on the number of factors to be extracted under different situations.

All of these lines of development may extend the applicability of factor analysis methods to a wider range of social research problems than has as yet been covered. Factor analysis in the future may prove as applicable for analyzing the interrelationships among manifestations on the part of groups as for the interrelationships among abilities of individuals.

SUGGESTED READINGS

Burt, Cyril, *The Factors of the Mind* (New York: Macmillan, 1941).

Cattell, Raymond B., *Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist* (New York: Harper, 1952).

Holzinger, Karl J., and Harman, Harry H., *Factor Analysis: A Synthesis of Factorial Methods* (Chicago: University of Chicago Press, 1941).

Horst, Paul, Wallin, Paul, Guttman, Louis, and others, *The Prediction of Personal Adjustment* (New York: Social Science Research Council, Bulletin 48, 1941).

Hotelling, Harold, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, pp. 417–441, 498–520.

Price, Daniel O., "Factor Analysis in the Study of Metropolitan Centers," *Social Forces*, 20 (May 1942), pp. 449–455.

Stephenson, W., "The Inverted Factor Technique," *British Journal of Psychology*, 26 (1936), pp. 344–361.

Thurstone, L. L., *Multiple-Factor Analysis: A Development and Expansion of The Vectors of Mind* (Chicago: University of Chicago Press, 1947).

Wilks, S. S., "Weighting Systems for Linear Functions of Correlated Variables When There is No Independent Variable," *Psychometrika*, 3 (March 1938), pp. 24–43.

Winch, Robert F., "Heuristic and Empirical Typologies: A Job for Factor Analysis," *American Sociological Review*, 12 (February 1947), pp. 68–75.

Wolfe, Dael, *Factor Analysis to 1940*, "Psychometric Monographs," No. 3 (Chicago: University of Chicago Press, 1940).

---

[16] *Ibid.*

Appendix

Most Frequently Used Formulas

$$p_i = l + \frac{\frac{jN}{100} - F}{f} \, i \qquad\qquad 120$$

$$s = \sqrt{\frac{\Sigma x^2}{N}} \qquad\qquad 121$$

$$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}} \qquad\qquad 121$$

$$s = \frac{1}{N}\sqrt{N\Sigma X^2 - (\Sigma X)^2} \qquad\qquad 121$$

$$s = i\sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \qquad\qquad 124$$

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \qquad\qquad 124$$

$$z = \frac{X - \bar{X}}{s} = \frac{x}{s} \qquad\qquad 125$$

$$V = \frac{s}{\bar{X}} \qquad\qquad 126$$

Measures of form and related formulas

$$Y_c = \frac{Ni}{s\sqrt{2\pi}} \, e^{\frac{-(x - \bar{x})^2}{2s^2}} \qquad\qquad 207$$

$$Y_o = .39894 \, \frac{Ni}{s} \qquad\qquad 209$$

$$\mu_1 = \frac{\Sigma x}{N} = 0 \qquad\qquad 210$$

$$\mu_2 = \frac{\Sigma x^2}{N} \qquad\qquad 210$$

$$\mu_3 = \frac{\Sigma x^3}{N} \qquad\qquad 211$$

$$\mu_4 = \frac{\Sigma x^4}{N} \qquad\qquad 211$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad\qquad 211$$

Formula

$$\chi^2 = \sum \frac{(f - f_c)^2}{f_c}$$

267, 365

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

370

$$\phi^2 = \frac{\chi^2}{N}$$

370

$$T^2 = \frac{\phi^2}{\sqrt{(s - 1)(t - 1)}}$$

371

Formulas associated with analysis of variance

Between-class sum of squares $= \dfrac{(X_{.1})^2}{k_1} + \dfrac{(X_{.2})^2}{k_2} + \cdots + \dfrac{(X_{.m})^2}{k_m} - \dfrac{(X_{..})^2}{N}$

402

$$\text{Intraclass } r = \frac{V_b - V_w}{V_b + (N - 1)V_w}$$

392

$$\epsilon^2 = 1 - \frac{V_w}{V_t}$$

393

Correlation, regression, and related formulas

$$\Sigma xy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}$$

412

$$r_{XY} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

412

$$r_{XY} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

413

$$Y_{c.x} = a_{YX} + b_{YX}X$$

415

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

416

$$a_{YX} = \frac{\Sigma Y - b\Sigma X}{N}$$

416

$$\sigma_r = \frac{1}{\sqrt{N - 1}} \text{ (when } \rho = 0)$$

424

$$\sigma_{\hat{z}} = \frac{1}{\sqrt{N - 3}}$$

428

$$F = \frac{r^2(N - 2)}{1 - r^2}$$

431

Table A. AREAS UNDER THE NORMAL CURVE

Fractional parts of the total area (10,000) under the normal probability curve, corresponding to the distances on the baseline between the mean and successive points of division laid off from the mean. Distances are measured in units of the standard deviation, σ. To illustrate, the table is read as follows: between the mean ordinate, y_o, and any ordinate erected at a distance from it of, say, $.8\sigma$ $\left(i.e., \dfrac{x}{\sigma} = .8 \right)$ is included 28.81 percent of the entire area.

$\dfrac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0.7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3718	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4083	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1.6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4758	4762	4767
2.0	4773	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2.9	4981	4982	4983	4984	4984	4984	4985	4985	4986	4986
3.0	4986.5	4987	4987	4988	4988	4988	4989	4989	4989	4990
3.1	4990.0	4991	4991	4991	4992	4992	4992	4992	4993	4993
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

Source: Harold O. Rugg, *Statistical Methods Applied to Education* (New York: Houghton Mifflin Company, 1917), Appendix Table III, pp. 389–390.

Table B. ORDINATES OF THE NORMAL CURVE

Ordinates of the normal probability curve expressed as fractional parts of the mean ordinate y_0. Each ordinate is erected at a given distance from the mean. The height of the ordinate erected at the mean can be computed from,

$$y_0 = \frac{N}{\sigma \sqrt{2 \pi}} = \frac{N}{2.5066\, \sigma}$$

The corresponding height of any other ordinate can be read from the table by assigning the distance that the ordinate is from the mean (x). Distances on x are measured as fractional parts of σ. Thus the height of an ordinate at a distance from the mean of $.7\sigma$ will be $.78270\ y_0$; the height of an ordinate at $2.15\ \sigma$ from the mean will be $.09914\ y_0$, etc.

x/σ	0	1	2	3	4	5	6	7	8	9
0.0	100000	99995	99980	99955	99920	99875	99820	99755	99685	99596
0.1	99501	99396	99283	99158	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96420	96156	95882
0.3	95600	95309	95010	94702	94387	94055	93723	93382	93024	92677
0.4	92312	91939 *	91558	91169	90774	90371	89961	89543	89119	88688
0.5	88250	87805	87353	86896	86432	85962	85488	85006	84519	84060
0.6	83527	83023	82514	82010	81481	80957	80429	79896	79359	78817
0.7	78270	77721	77167	76610	76048	75484	74916	74342	73769	73193
0.8	72615	72033	71448	70861	70272	69681	69087	68493	67896	67298
0.9	66689	66097	65494	64891	64287	63683	63077	62472	61865	61259
1.0	60653	60047	59440	58834	58228	57623	57017	56414	55810	55209
1.1	54607	54007	53409	52812	52214	51620	51027	50437	49848	49260
1.2	48675	48092	47511	46933	46357	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38569	38058
1.4	37531	37007	36487	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31500	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22008	21627	21251	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1.9	16448	16137	15831	15530	15232	14939	14650	14364	14083	13806
2.0	13534	13265	13000	12740	12483	12230	11981	11737	11496	11259
2.1	11025	10795	10570	10347	10129	09914	09702	09495	09290	09090
2.2	08892	08698	08507	08320	08136	07956	07778	07604	07433	07265
2.3	07100	06939	06780	06624	06471	06321	06174	06029	05888	05750
2.4	05614	05481	05350	05222	05096	04973	04852	04734	04618	04505
2.5	04394	04285	04179	04074	03972	03873	03775	03680	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02684
2.7	02612	02542	02474	02408	02343	02280	02218	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01288	01252	01215	01179	01145
3.0	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.0	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.0	00000									

* Value corrected from that appearing in original table.

MEAN

$\dfrac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	1.0000	.9920	.9840	.9760	.9682	.9602	.9522	.9442	.9362	.9282
0.1	.9204	.9124	.9044	.8966	.8886	.8808	.8728	.8650	.8572	.8492
0.2	.8414	.8336	.8258	.8180	.8104	.8026	.7948	.7872	.7794	.7718
0.3	.7642	.7566	.7490	.7414	.7338	.7264	.7188	.7114	.7040	.6966
0.4	.6892	.6818	.6744	.6672	.6600	.6528	.6456	.6384	.6312	.6242
0.5	.6170	.6100	.6030	.5962	.5892	.5824	.5754	.5686	.5612	.5552
0.6	.5486	.5418	.5352	.5286	.5222	.5156	.5092	.5028	.4964	.4902
0.7	.4840	.4776	.4716	.4654	.4592	.4532	.4472	.4412	.4354	.4296
0.8	.4238	.4180	.4122	.4066	.4010	.3954	.3898	.3844	.3788	.3734
0.9	.3682	.3628	.3576	.3524	.3472	.3422	.3370	.3320	.3270	.3222
1.0	.3174	.3124	.3078	.3030	.2984	.2938	.2892	.2846	.2802	.2758
1.1	.2714	.2670	.2628	.2564	.2542	.2502	.2460	.2420	.2380	.2338
1.2	.2302	.2262	.2224	.2186	.2150	.2112	.2076	.2040	.2006	.1970
1.3	.1936	.1902	.1868	.1834	.1802	.1770	.1738	.1706	.1676	.1646
1.4	.1616	.1586	.1556	.1528	.1498	.1470	.1442	.1416	.1388	.1362
1.5	.1336	.1310	.1286	.1260	.1236	.1212	.1188	.1164	.1140	.1118
1.6	.1096	.1074	.1052	.1030	.1010	.0990	.0970	.0950	.0930	.0910
1.7	.0892	.0872	.0854	.0836	.0818	.0802	.0784	.0768	.0750	.0734
1.8	.0718	.0702	.0688	.0672	.0658	.0644	.0628	.0614	.0602	.0588
1.9	.0574	.0562	.0548	.0536	.0524	.0512	.0500	.0484	.0476	.0466
2.0	.0454	.0444	.0434	.0424	.0414	.0404	.0394	.0384	.0376	.0366
2.1	.0358	.0348	.0340	.0332	.0324	.0316	.0308	.0300	.0292	.0286
2.2	.0278	.0270	.0264	.0258	.0250	.0244	.0238	.0232	.0226	.0220
2.3	.0214	.0208	.0204	.0198	.0192	.0188	.0182	.0178	.0174	.0168
2.4	.0164	.0160	.0156	.0150	.0146	.0142	.0138	.0136	.0132	.0128
2.5	.0124	.0120	.0118	.0114	.0110	.0108	.0104	.0102	.0098	.0096
2.6	.0094	.0090	.0088	.0086	.0082	.0080	.0078	.0076	.0074	.0072
2.7	.0070	.0068	.0066	.0064	.0062	.0060	.0058	.0056	.0054	.0052
2.8	.0052	.0050	.0048	.0046	.0046	.0044	.0042	.0040	.0040	.0038
2.9	.0038	.0036	.0034	.0032	.0032	.0032	.0030	.0030	.0028	.0028
3.0	.0027	.0026	.0026	.0024	.0024	.0024	.0022	.0022	.0022	.0020
3.1	.0020	.0018	.0018	.0018	.0016	.0016	.0016	.0016	.0014	.0014
3.2	.00137									
3.3	.0009668									
3.4	.0006738									
3.5	.0004652									
3.6	.0003182									
3.7	.0002156									
3.8	.0001446									
3.9	.0000962									
4.0	.0000634									
4.5	.0000068									
5.0	.00000057									

Source: Table A.

Table D. DISTRIBUTION OF t
Probability

n	·9	·8	·7	·6	·5	·4	·3	·2	·1	·05	·02	·01	·001
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657	636·619
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925	31·598
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841	12·941
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604	8·610
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032	6·859
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707	5·959
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499	5·405
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355	5·041
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250	4·781
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169	4·587
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106	4·437
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055	4·318
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012	4·221
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977	4·140
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947	4·073
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921	4·015
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898	3·965
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878	3·922
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861	3·883
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845	3·850
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831	3·819
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819	3·792
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807	3·767
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797	3·745
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787	3·725
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779	3·707
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771	3·690
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763	3·674
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756	3·659
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750	3·646
40	·126	·255	·388	·529	·681	·851	1·050	1·303	1·684	2·021	2·423	2·704	3·551
60	·126	·254	·387	·527	·679	·848	1·046	1·296	1·671	2·000	2·390	2·660	3·460
120	·126	·254	·386	·526	·677	·845	1·041	1·289	1·658	1·980	2·358	2·617	3·373
∞	·126	·253	·385	·524	·674	·842	1·036	1·282	1·645	1·960	2·326	2·576	3·291

Table E. DISTRIBUTION OF χ^2
Probability

n	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	$.0^3157$	$.0^2628$	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341	16.268
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{n^2 - 1}$ may be used as a normal deviate with unit variance.

Source: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (London: Oliver and Boyd, 1938), Table IV, p. 27.

Table F. VARIANCE RATIO (F)
5-Percent Points of e^{2z}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	238·9	243·9	249·0	254·3
2	18·51	19·00	19·16	19·25	19·30	19·33	19·37	19·41	19·45	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·84	8·74	8·64	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·04	5·91	5·77	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·82	4·68	4·53	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·15	4·00	3·84	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·73	3·57	3·41	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·44	3·28	3·12	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·23	3·07	2·90	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·07	2·91	2·74	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	2·95	2·79	2·61	2·40
12	4·75	3·88	3·49	3·26	3·11	3·00	2·85	2·69	2·50	2·30
13	4·67	3·80	3·41	3·18	3·02	2·92	2·77	2·60	2·42	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·70	2·53	2·35	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·64	2·48	2·29	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·59	2·42	2·24	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·55	2·38	2·19	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·51	2·34	2·15	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·48	2·31	2·11	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·45	2·28	2·08	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·42	2·25	2·05	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·40	2·23	2·03	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·38	2·20	2·00	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·36	2·18	1·98	1·73
25	4·24	3·38	2·99	2·76	2·60	2·49	2·34	2·16	1·96	1·71
26	4·22	3·37	2·98	2·74	2·59	2·47	2·32	2·15	1·95	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·30	2·13	1·93	1·67
28	4·20	3·34	2·95	2·71	2·56	2·44	2·29	2·12	1·91	1·65
29	4·18	3·33	2·93	2·70	2·54	2·43	2·28	2·10	1·90	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·27	2·09	1·89	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·18	2·00	1·79	1·51
60	4·00	3·15	2·76	2·52	2·37	2·25	2·10	1·92	1·70	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·02	1·83	1·61	1·25
∞	3·84	2·99	2·60	2·37	2·21	2·09	1·94	1·75	1·52	1·00

Lower 5-percent points are found by interchange of n_1 and n_2, i.e. n_1 must always correspond with the greater mean square.

Source: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (London: Oliver and Boyd, 1938), pp. 31, 33, 35, of Table V.

1-Percent Points of e^{2z}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98·49	99·01	99·17	99·25	99·30	99·33	99·36	99·42	99·46	99·50
3	34·12	30·81	29·46	28·71	28·24	27·91	27·49	27·05	26·60	26·12
4	21·20	18·00	16·69	15·98	15·52	15·21	14·80	14·37	13·93	13·46
5	16·26	13·27	12·06	11·39	10·97	10·67	10·27	9·89	9·47	9·02
6	13·74	10·92	9·78	9·15	8·75	8·47	8·10	7·72	7·31	6·88
7	12·25	9·55	8·45	7·85	7·46	7·19	6·84	6·47	6·07	5·65
8	11·26	8·65	7·59	7·01	6·63	6·37	6·03	5·67	5·28	4·86
9	10·56	8·02	6·99	6·42	6·06	5·80	5·47	5·11	4·73	4·31
10	10·04	7·56	6·55	5·99	5·64	5·39	5·06	4·71	4·33	3·91
11	9·65	7·20	6·22	5·67	5·32	5·07	4·74	4·40	4·02	3·60
12	9·33	6·93	5·95	5·41	5·06	4·82	4·50	4·16	3·78	3·36
13	9·07	6·70	5·74	5·20	4·86	4·62	4·30	3·96	3·59	3·16
14	8·86	6·51	5·56	5·03	4·69	4·46	4·14	3·80	3·43	3·00
15	8·68	6·36	5·42	4·89	4·56	4·32	4·00	3·67	3·29	2·87
16	8·53	6·23	5·29	4·77	4·44	4·20	3·89	3·55	3·18	2·75
17	8·40	6·11	5·18	4·67	4·34	4·10	3·79	3·45	3·08	2·65
18	8·28	6·01	5·09	4·58	4·25	4·01	3·71	3·37	3·00	2·57
19	8·18	5·93	5·01	4·50	4·17	3·94	3·63	3·30	2·92	2·49
20	8·10	5·85	4·94	4·43	4·10	3·87	3·56	3·23	2·86	2·42
21	8·02	5·78	4·87	4·37	4·04	3·81	3·51	3·17	2·80	2·36
22	7·94	5·72	4·82	4·31	3·99	3·76	3·45	3·12	2·75	2·31
23	7·88	5·66	4·76	4·26	3·94	3·71	3·41	3·07	2·70	2·26
24	7·82	5·61	4·72	4·22	3·90	3·67	3·36	3·03	2·66	2·21
25	7·77	5·57	4·68	4·18	3·86	3·63	3·32	2·99	2·62	2·17
26	7·72	5·53	4·64	4·14	3·82	3·59	3·29	2·96	2·58	2·13
27	7·68	5·49	4·60	4·11	3·78	3·56	3·26	2·93	2·55	2·10
28	7·64	5·45	4·57	4·07	3·75	3·53	3·23	2·90	2·52	2·06
29	7·60	5·42	4·54	4·04	3·73	3·50	3·20	2·87	2·49	2·03
30	7·56	5·39	4·51	4·02	3·70	3·47	3·17	2·84	2·47	2·01
40	7·31	5·18	4·31	3·83	3·51	3·29	2·99	2·66	2·29	1·80
60	7·08	4·98	4·13	3·65	3·34	3·12	2·82	2·50	2·12	1·60
120	6·85	4·79	3·95	3·48	3·17	2·96	2·66	2·34	1·95	1·38
∞	6·64	4·60	3·78	3·32	3·02	2·80	2·51	2·18	1·79	1·00

Lower 1-percent points are found by interchange of n_1 and n_2, *i.e.* n_1 must always correspond with the greater mean square.

0·1-Percent Points of e^{2z}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	405284	500000	540379	562500	576405	585937	598144	610667	623497	636619
2	998·5	999·0	999·2	999·2	999·3	999·3	999·4	999·4	999·5	999·5
3	167·5	148·5	141·1	137·1	134·6	132·8	130·6	128·3	125·9	123·5
4	74·14	61·25	56·18	53·44	51·71	50·53	49·00	47·41	45·77	44·05
5	47·04	36·61	33·20	31·09	29·75	28·84	27·64	26·42	25·14	23·78
6	35·51	27·00	23·70	21·90	20·81	20·03	19·03	17·99	16·89	15·75
7	29·22	21·69	18·77	17·19	16·21	15·52	14·63	13·71	12·73	11·69
8	25·42	18·49	15·83	14·39	13·49	12·86	12·04	11·19	10·30	9·34
9	22·86	16·39	13·90	12·56	11·71	11·13	10·37	9·57	8·72	7·81
10	21·04	14·91	12·55	11·28	10·48	9·92	9·20	8·45	7·64	6·76
11	19·69	13·81	11·56	10·35	9·58	9·05	8·35	7·63	6·85	6·00
12	18·64	12·97	10·80	9·63	8·89	8·38	7·71	7·00	6·25	5·42
13	17·81	12·31	10·21	9·07	8·35	7·86	7·21	6·52	5·78	4·97
14	17·14	11·78	9·73	8·62	7·92	7·43	6·80	6·13	5·41	4·60
15	16·59	11·34	9·34	8·25	7·57	7·09	6·47	5·81	5·10	4·31
16	16·12	10·97	9·00	7·94	7·27	6·81	6·19	5·55	4·85	4·06
17	15·72	10·66	8·73	7·68	7·02	6·56	5·96	5·32	4·63	3·85
18	15·38	10·39	8·49	7·46	6·81	6·35	5·76	5·13	4·45	3·67
19	15·08	10·16	8·28	7·26	6·61	6·18	5·59	4·97	4·29	3·52
20	14·82	9·95	8·10	7·10	6·46	6·02	5·44	4·82	4·15	3·38
21	14·59	9·77	7·94	6·95	6·32	5·88	5·31	4·70	4·03	3·26
22	14·38	9·61	7·80	6·81	6·19	5·76	5·19	4·58	3·92	3·15
23	14·19	9·47	7·67	6·69	6·08	5·65	5·09	4·48	3·82	3·05
24	14·03	9·34	7·55	6·59	5·98	5·55	4·99	4·39	3·74	2·97
25	13·88	9·22	7·45	6·49	5·88	5·46	4·91	4·31	3·66	2·89
26	13·74	9·12	7·36	6·41	5·80	5·38	4·83	4·24	3·59	2·82
27	13·61	9·02	7·27	6·33	5·73	5·31	4·76	4·17	3·52	2·75
28	13·50	8·93	7·19	6·25	5·66	5·24	4·69	4·11	3·46	2·70
29	13·39	8·85	7·12	6·19	5·59	5·18	4·64	4·05	3·41	2·64
30	13·29	8·77	7·05	6·12	5·53	5·12	4·58	4·00	3·36	2·59
40	12·61	8·25	6·60	5·70	5·13	4·73	4·21	3·64	3·01	2·23
60	11·97	7·76	6·17	5·31	4·76	4·37	3·87	3·31	2·69	1·90
120	11·38	7·31	5·79	4·95	4·42	4·04	3·55	3·02	2·40	1·56
∞	10·83	6·91	5·42	4·62	4·10	3·74	3·27	2·74	2·13	1·00

Lower 0·1-percent points are found by interchange of n_1 and n_2, *i.e.* n_1 must always correspond with the greater mean square.

$$Z = \frac{1}{2} \log_e \frac{1+r}{1-r}$$

Computed by evaluating $Z = 1.151293 \, [\log_{10} (1 + r) - \log_{10} (1 - r)]$ with logarithms taken from C. Bruhns, *Neues logarithmisch-trigonometrisches Handbuch auf sieben Decimalen* (Leipzig: Verlag Von Bernhard Tauchnitz, 1913), Table 1.

r	0	1	2	3	4	5	6	7	8	9
.00	.000000	.001000	.002000	.003000	.004000	.005000	.006000	.007000	.008000	.009000
.01	.010000	.011000	.012001	.013001	.014001	.015001	.016001	.017002	.018002	.019002
.02	.020003	.021003	.022004	.023004	.024005	.025005	.026006	.027007	.028007	.029008
.03	.030009	.031010	.032011	.033012	.034013	.035014	.036016	.037017	.038018	.039020
.04	.040021	.041023	.042025	.043027	.044028	.045030	.046032	.047035	.048037	.049039
.05	.050042	.051044	.052047	.053050	.054053	.055056	.056059	.057062	.058065	.059069
.06	.060072	.061076	.062080	.063084	.064088	.065092	.066096	.067101	.068105	.069110
.07	.070115	.071120	.072125	.073130	.074136	.075141	.076147	.077153	.078159	.079165
.08	.080171	.081178	.082185	.083192	.084198	.085206	.086213	.087220	.088228	.089236
.09	.090244	.091253	.092261	.093270	.094278	.095287	.096297	.097306	.098316	.099325
.10	.100335	.101346	.102356	.103367	.104377	.105389	.106400	.107411	.108423	.109435
.11	.110447	.111459	.112472	.113485	.114498	.115511	.116525	.117538	.118552	.119567
.12	.120581	.121596	.122611	.123626	.124642	.125657	.126673	.127690	.128706	.129723
.13	.130740	.131757	.132775	.133793	.134811	.135829	.136848	.137867	.138886	.139906
.14	.140926	.141946	.142966	.143987	.145008	.146029	.147051	.148073	.149095	.150118
.15	.151141	.152164	.153187	.154211	.155235	.156260	.157284	.158309	.159335	.160361
.16	.161387	.162413	.163440	.164467	.165495	.166522	.167551	.168579	.169608	.170637
.17	.171667	.172697	.173727	.174758	.175789	.176820	.177852	.178884	.179917	.180949
.18	.181983	.183016	.184051	.185085	.186120	.187155	.188191	.189227	.190263	.191300
.19	.192337	.193375	.194413	.195451	.196490	.197530	.198569	.199610	.200650	.201691
.20	.202733	.203774	.204817	.205860	.206903	.207946	.208991	.210035	.211080	.212126
.21	.213171	.214218	.215265	.216312	.217360	.218408	.219457	.220506	.221555	.222606
.22	.223656	.224707	.225759	.226811	.227864	.228917	.229970	.231024	.232079	.233134
.23	.234190	.235246	.236302	.237359	.238417	.239475	.240534	.241593	.242653	.243713
.24	.244774	.245836	.246898	.247960	.249023	.250087	.251151	.252215	.253281	.254347
.25	.255413	.256480	.257547	.258616	.259684	.260753	.261823	.262894	.263965	.265036
.26	.266108	.267181	.268255	.269329	.270403	.271479	.272554	.273631	.274708	.275786
.27	.276864	.277943	.279022	.280103	.281184	.282265	.283347	.284430	.285513	.286597
.28	.287682	.288768	.289854	.290940	.292028	.293116	.294205	.295294	.296384	.297475
.29	.298566	.299659	.300751	.301845	.302939	.304034	.305130	.306226	.307323	.308421
.30	.309520	.310625	.311719	.312820	.313921	.315023	.316126	.317230	.318334	.319440
.31	.320546	.321652	.322760	.323868	.324977	.326087	.327197	.328309	.329421	.330534
.32	.331647	.332762	.333877	.334993	.336110	.337228	.338346	.339465	.340586	.341707
.33	.342828	.343951	.345074	.346199	.347324	.348450	.349577	.350704	.351833	.352962
.34	.354093	.355224	.356356	.357489	.358623	.359757	.360893	.362029	.363167	.364305
.35	.365444	.366584	.367725	.368867	.370010	.371153	.372298	.373444	.374590	.375738
.36	.376886	.378035	.379186	.380337	.381489	.382643	.383797	.384952	.386108	.387265
.37	.388423	.389582	.390743	.391904	.393066	.394229	.395393	.396556	.397724	.398892
.38	.400060	.401229	.402399	.403571	.404743	.405917	.407091	.408267	.409444	.410621
.39	.411800	.412980	.414161	.415343	.416527	.417711	.418896	.420083	.421270	.422459
.40	.423649	.424840	.426032	.427226	.428420	.429616	.430813	.432010	.433210	.434410
.41	.435611	.436814	.438018	.439223	.440429	.441637	.442845	.444055	.445266	.446479
.42	.447692	.448907	.450123	.451340	.452559	.453779	.455000	.456222	.457446	.458671
.43	.459897	.461124	.462353	.463583	.464815	.466047	.467281	.468517	.469754	.470992
.44	.472231	.473472	.474714	.475957	.477202	.478448	.479696	.480945	.482195	.483447
.45	.484700	.485955	.487211	.488469	.489728	.490988	.492250	.493513	.494778	.496044
.46	.497311	.498581	.499851	.501123	.502397	.503672	.504949	.506227	.507507	.508788
.47	.510070	.511355	.512641	.513928	.515217	.516508	.517800	.519094	.520389	.521686
.48	.522985	.524285	.525586	.526890	.528195	.529502	.530810	.532120	.533432	.534745
.49	.536061	.537377	.538696	.540016	.541338	.542662	.543987	.545314	.546643	.547974
.50	.549306	.550641	.551977	.553314	.554654	.555995	.557339	.558684	.560031	.561379

$$Z = \frac{1}{2} \log_e \frac{1+r}{1-r}$$

Computed by evaluating $Z = 1.151293 \, [\log_{10} (1 + r) - \log_{10} (1 - r)]$ with logarithms taken from C. Bruhns, *Neues logarithmisch-trigonometrisches Handbuch auf sieben Decimalen* (Leipzig: Verlag Von Bernhard Tauchnitz, 1913), Table 1.

r	0	1	2	3	4	5	6	7	8	9
.51	.562730	.564082	.565437	.566793	.568151	.569511	.570873	.572237	.573603	.574970
.52	.576340	.577712	.579085	.580461	.581838	.583218	.584599	.585983	.587368	.588756
.53	.590145	.591537	.592931	.594327	.595724	.597124	.598526	.599931	.601337	.602745
.54	.604156	.605568	.606983	.608400	.609820	.611241	.612665	.614091	.615519	.616949
.55	.618382	.619816	.621253	.622693	.624134	.625579	.627025	.628473	.629924	.631378
.56	.632833	.634292	.635752	.637215	.638680	.640148	.641618	.643091	.644566	.646043
.57	.647523	.649006	.650491	.651978	.653468	.654961	.656456	.657954	.659454	.660957
.58	.662463	.663971	.665482	.666996	.668512	.670031	.671553	.673077	.674604	.676134
.59	.677666	.679202	.680740	.682281	.683825	.685371	.686921	.688473	.690028	.691586
.60	.693147	.694711	.696278	.697848	.699421	.700997	.702576	.704158	.705742	.707330
.61	.708922	.710516	.712113	.713713	.715317	.716924	.718534	.720147	.721763	.723382
.62	.725005	.726631	.728261	.729893	.731529	.733169	.734811	.736458	.738107	.739760
.63	.741416	.743076	.744740	.746406	.748077	.749751	.751428	.753109	.754794	.756482
.64	.758174	.759870	.761569	.763272	.764979	.766689	.768403	.770122	.771844	.773569
.65	.775299	.777033	.778770	.780511	.782257	.784006	.785760	.787517	.789279	.791044
.66	.792814	.794588	.796366	.798148	.799934	.801725	.803520	.805320	.807123	.808931
.67	.810744	.812560	.814381	.816207	.818037	.819872	.821711	.823555	.825404	.827257
.68	.829114	.830977	.832844	.834716	.836593	.838474	.840361	.842252	.844149	.846050
.69	.847956	.849867	.851784	.853705	.855632	.857564	.859500	.861443	.863390	.865343
.70	.867300	.869264	.871233	.873208	.875187	.877173	.879164	.881160	.883163	.885171
.71	.887184	.889204	.891229	.893260	.895297	.897340	.899389	.901444	.903505	.905572
.72	.907645	.909725	.911811	.913903	.916001	.918106	.920217	.922335	.924459	.926590
.73	.928728	.930872	.933023	.935181	.937345	.939517	.941695	.943881	.946073	.948273
.74	.950480	.952694	.954915	.957144	.959380	.961623	.963875	.966133	.968400	.970674
.75	.972955	.975245	.977543	.979848	.982162	.984483	.986813	.989151	.991498	.993852
.76	.996216	.998587	1.000967	1.003356	1.005754	1.008161	1.010576	1.013000	1.015434	1.017876
.77	1.020328	1.022789	1.025260	1.027739	1.030229	1.032728	1.035237	1.037755	1.040284	1.042822
.78	1.045371	1.047930	1.050499	1.053078	1.055668	1.058268	1.060879	1.063501	1.066134	1.068777
.79	1.071432	1.074098	1.076775	1.079579	1.082164	1.084876	1.087599	1.090334	1.093082	1.095841
.80	1.098613	1.101397	1.104193	1.107002	1.109824	1.112659	1.115506	1.118367	1.121241	1.124128
.81	1.127030	1.129944	1.132873	1.135815	1.138772	1.141743	1.144728	1.147728	1.150743	1.153773
.82	1.156818	1.159878	1.162954	1.166045	1.169152	1.172275	1.175414	1.178570	1.181742	1.184931
.83	1.188137	1.191360	1.194600	1.197858	1.201134	1.204428	1.207740	1.211070	1.214419	1.217787
.84	1.221174	1.224581	1.228007	1.231452	1.234919	1.238405	1.241912	1.245441	1.248990	1.252561
.85	1.256153	1.259768	1.263405	1.267065	1.270747	1.274454	1.278183	1.281937	1.285715	1.289518
.86	1.293345	1.297198	1.301077	1.304982	1.308913	1.312871	1.316857	1.320870	1.324911	1.328981
.87	1.333080	1.337209	1.341367	1.345555	1.349775	1.354026	1.358308	1.362623	1.366971	1.371353
.88	1.375768	1.380218	1.384703	1.389224	1.393782	1.398376	1.403008	1.407678	1.412388	1.417137
.89	1.421926	1.426757	1.431630	1.436545	1.441504	1.446507	1.451556	1.456651	1.461792	1.466982
.90	1.472220	1.477508	1.482847	1.488239	1.493683	1.499181	1.504734	1.510344	1.516012	1.521738
.91	1.527525	1.533373	1.539285	1.545261	1.551302	1.557411	1.563590	1.569839	1.576160	1.582556
.92	1.589027	1.595577	1.602207	1.608913	1.615715	1.622597	1.629568	1.636631	1.643787	1.651039
.93	1.658391	1.665844	1.673402	1.681069	1.688846	1.696738	1.704749	1.712881	1.721139	1.729528
.94	1.738050	1.746711	1.755516	1.764469	1.773577	1.782843	1.792274	1.801877	1.811658	1.821624
.95	1.831782	1.842139	1.852705	1.863488	1.874497	1.885742	1.897234	1.908985	1.921005	1.933309
.96	1.945911	1.958825	1.972068	1.985656	1.999611	2.013951	2.028699	2.043880	2.059519	2.075648
.97	2.092296	2.109501	2.127300	2.145738	2.164861	2.184725	2.205389	2.226922	2.249400	2.272913
.98	2.297561	2.323460	2.350746	2.379577	2.410142	2.442663	2.477411	2.514717	2.554990	2.598747
.99	2.646653	2.699585	2.758727	2.825744	2.903070	2.994482	3.106304	3.250396	3.453379	3.800203
1.00	∞									

Index

Index
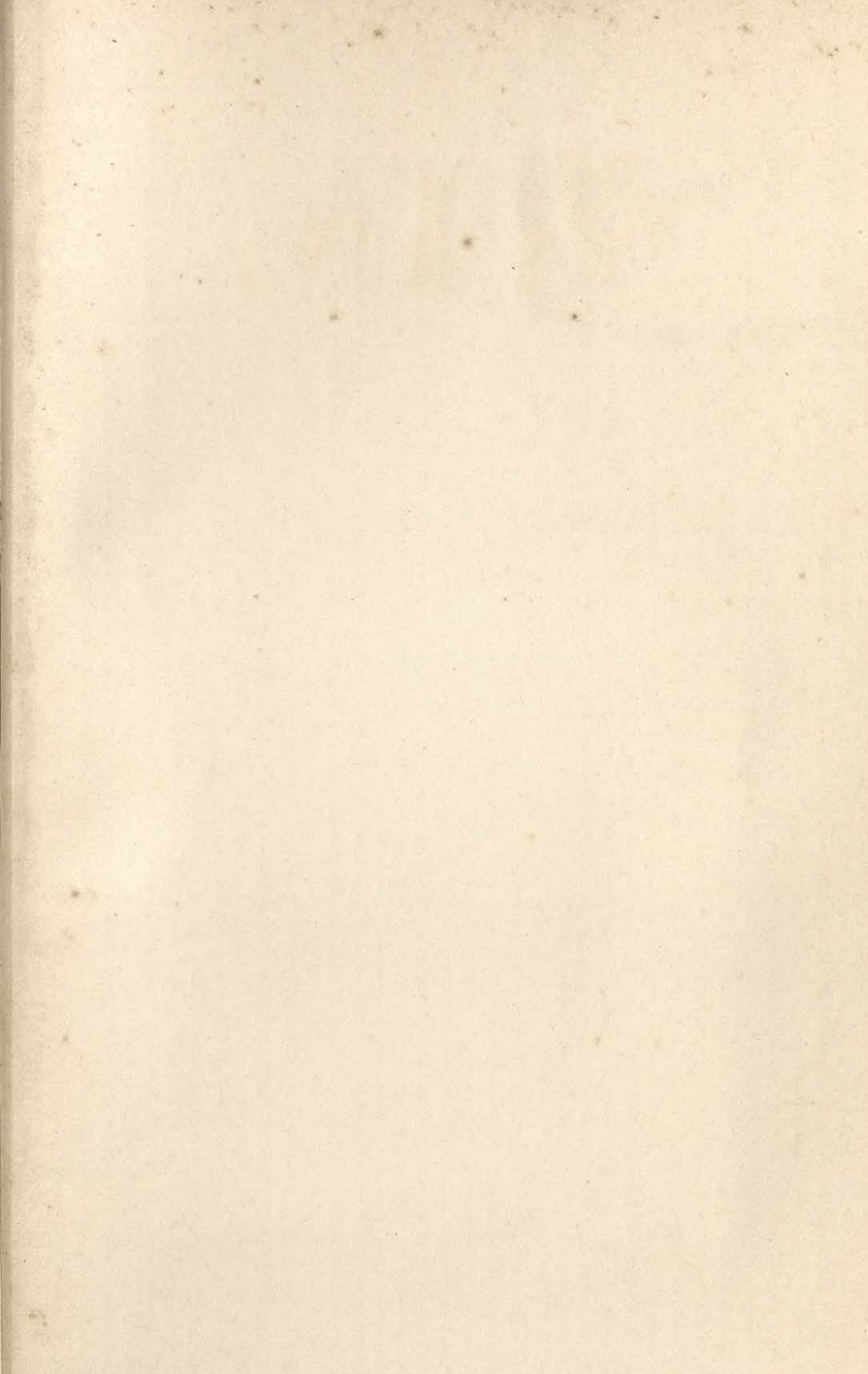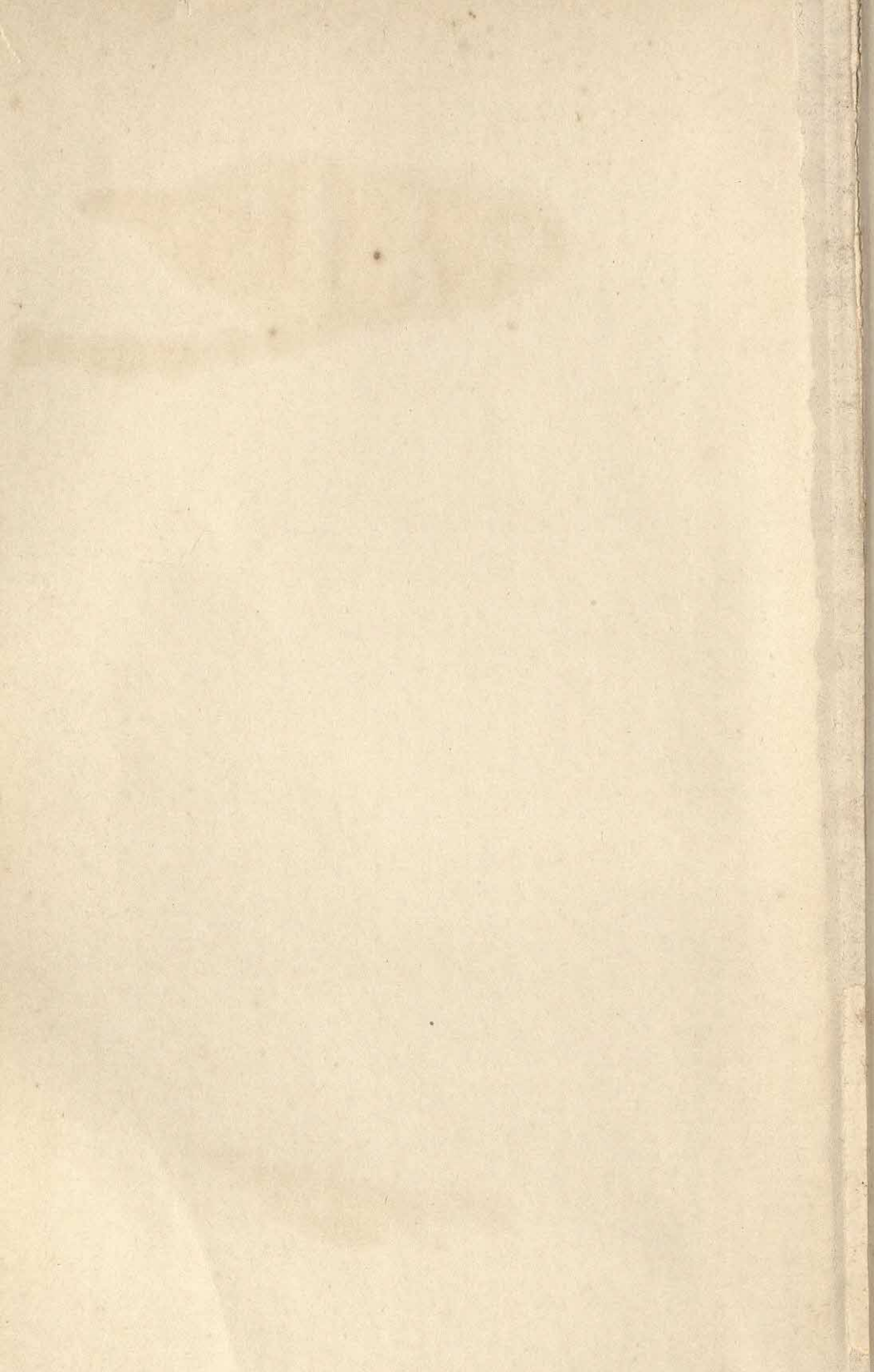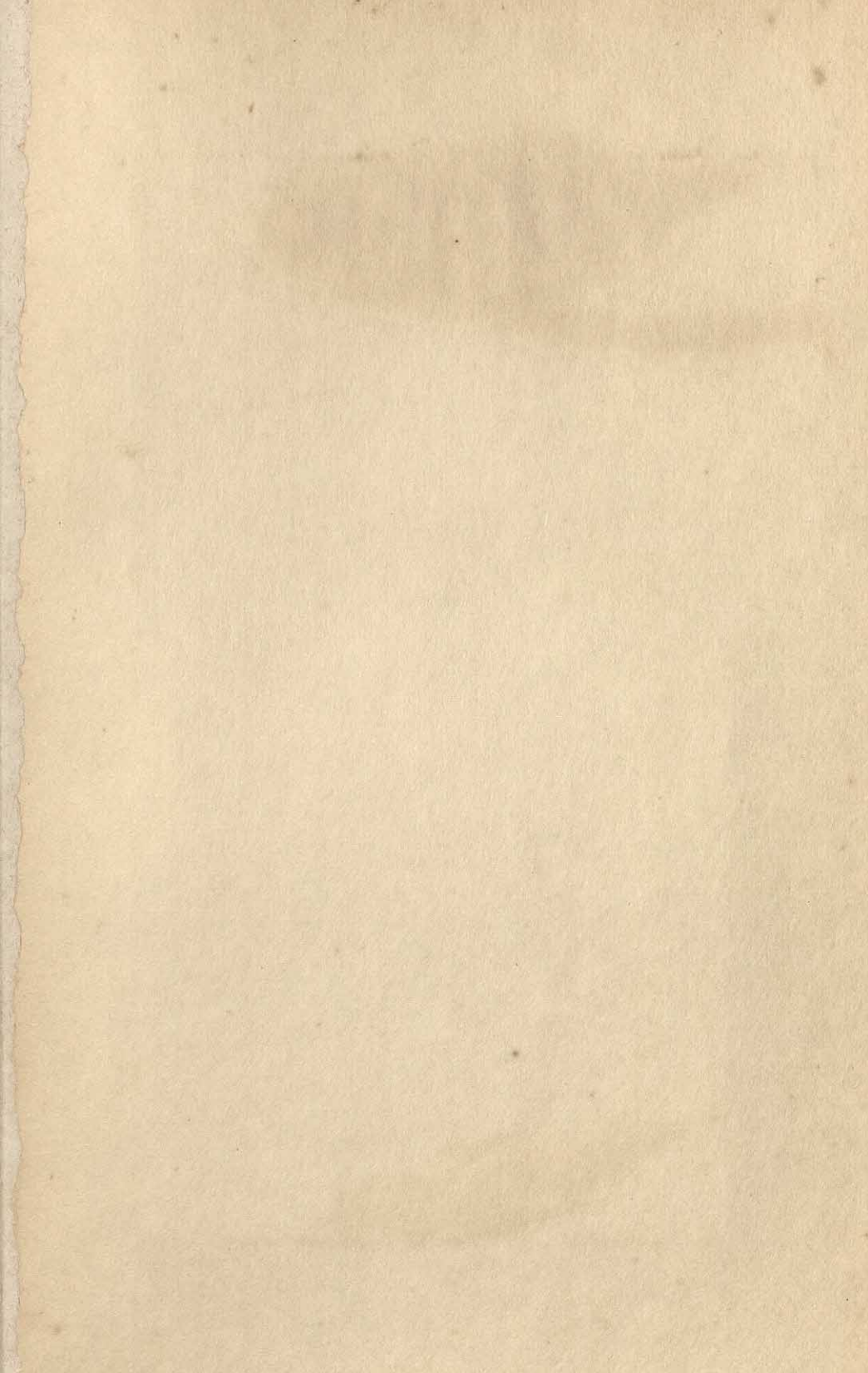
Form No. 3.

Bureau of Educational & Psychological Research Library.

The book is to be returned within the date stamped last.

21.12.72		
5.1.76		